User Facilities for 21st Century Systems Biology: Providing Critical Technologies for the Research Community

GENOMES to LIFE
BIOLOGICAL SOLUTIONS FOR ENERGY CHALLENGES
U.S. DEPARTMENT OF ENERGY
INNOVATIVE APPROACHES ALONG UNCONVENTIONAL PATHS

Office of Science
U.S. DEPARTMENT OF ENERGY

November 2002 DOEGenomesToLife.org

Executive Summary

Realizing the Potential of the Genome Revolution

The revolution in biology triggered by the Human Genome Project promises far-reaching benefits to our nation and environment. Today, scientists have in hand the complete DNA sequences of genomes for many organisms—from microbes to plants to humans. This knowledge makes it possible to address the ultimate goal of modern biology: to achieve a fundamental, comprehensive, and systematic understanding of life. This goal is founded, as is life itself, on the genome, which contains the basic information necessary for the construction and operation of a living organism. The new Genomes to Life (GTL) program combines advanced technologies with the information found in the DNA of microbial genomes to establish a foundation for achieving this goal. Obtaining a deep level of knowledge about the diverse natural capabilities of microbes will allow scientists, both in GTL and the broader scientific community, to use those capabilities to help solve challenges in energy security, environmental cleanup, and global climate change.

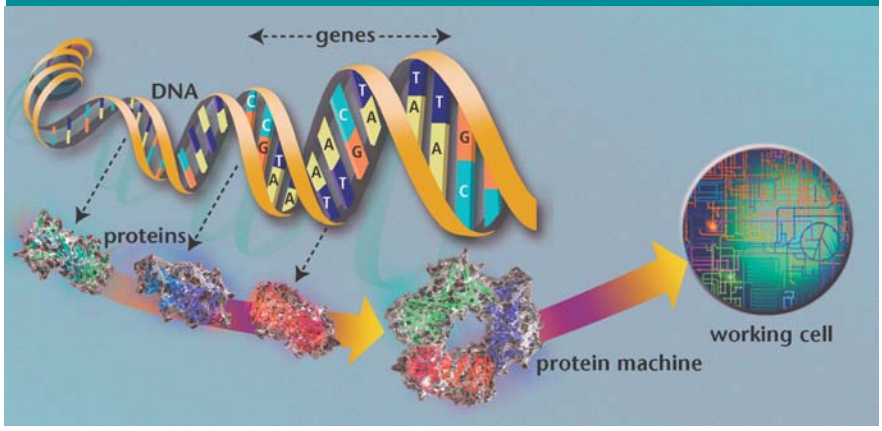
GTL scientific goals target the fundamental processes of living systems by studying them on three levels: (1) proteins and multicomponent molecular assemblies ("machines") that perform most of the cell's work, (2) gene regulatory networks that control these processes, and (3) microbial associations or communities in which groups of cells carry out the processes in nature. These tasks will require advanced experimental and computational methods and capabilities to assimilate, understand, and model the data on the scale and complexity of real living systems and, in the process, to build a dynamic knowledge base from this information. The resulting GTL knowledge base will provide data, models, and simulations of expression, pathways, and network systems, molecular machines, and cell and community processes for the entire research community.

Genomes to Life is designed to support the launching of biology onto a new trajectory—one that will empower biologists to pursue completely new approaches to discovery, based on explorations of whole functioning systems instead of the more traditional focus on cellular components. Genes encode proteins, which work together to carry out most of the activities in the cell (see figure below). These processes are orchestrated dynamically in an intricate labyrinth of pathways and networks that make the cell "come alive." Understanding cellular activities in a realistic context—known as systems biology—ultimately will transform biology to a more quantitative and predictive science that will enable effective and economical solutions to many of DOE's most pressing challenges. These capabilities will inspire revolutionary solutions to DOE mission challenges and transform the entire life sciences landscape, from agriculture to human health.

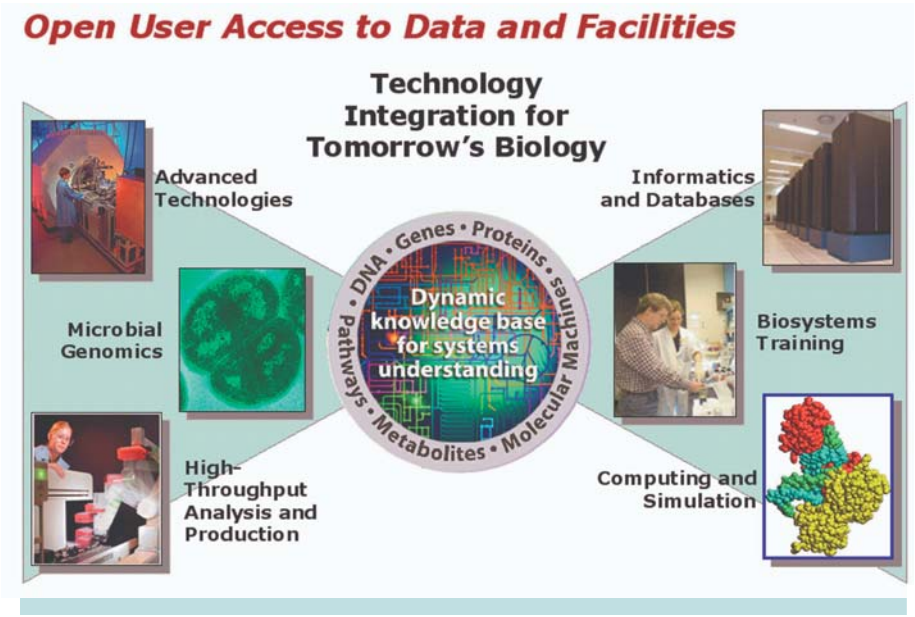
The Information Challenge

Thousands of microbes have capabilities of interest (see "Why Microbes," p. 4), and each microbial genome contains thousands of genes capable of

Genomes to Life : From DNA Sequence to Living Systems



GTL User Facilities Hallmarks



Four Enabling User Facilities

The technologies needed for systems biology require economies of scale achievable only at major facilities. To meet this challenge, the DOE Office of Biological and Environmental Research (BER) and Office of Advanced Scientific Computing Research (OASCR) propose to develop a powerful new core consisting of four complementary user facilities for both GTL and the broader scientific communities. Each will build on the capabilities of the others, moving from the use of genomic data to systematically identify, produce, and characterize microbial proteins and fully decipher how cells use the proteins to carry out life processes.

producing an even-greater number of proteins. Much biological research is now centered on completely characterizing proteins and their higher-order structures (sometimes called molecular machines) and relating that information to genome sequence. Molecular machines carry out chemical reactions, generate mechanical forces, transport metabolites and ions, and make possible every action of a biological system. Genomes contain regulatory elements that coordinate protein production and molecular machine assembly and are themselves cued by signals from the environment, including other microbial populations in their ecological community. A systems approach thus must extend from each genome throughout the population or community, encompassing thousands of proteins, molecular machines, pathways, networks, cells, and, eventually, their cellular systems and environments.

Just as DNA sequencing capability was completely inadequate at the beginning of the Human Genome Project, the quantity of data that must be collected and analyzed for systems-biology research far exceeds current capabilities and capacities. Collecting and using such data and reagents will require coordination and integration of dozens of high-throughput technologies and approaches, some not yet refined or even developed. The recent availability of genome data, emerging technologies, and high-performance computing and informatics tools and technologies now make such an approach practicable.

The facilities will make possible new avenues of inquiry, fundamentally changing the course of biological research and greatly accelerating the pace of discovery. Numerous piloting activities funded primarily by BER have established a foundation for future work and underscored the need for advanced technology user facilities accessible by the whole biological research community. Projects include systems biology and development of advanced technologies. To provide the powerful resources needed, the new facilities will use unique, high-throughput combinations of state-of-the-art instrumentation and technologies, automation, and tools. These time-phased facilities will be optimized and developed concurrently with research programs and integral computing and information infrastructure and tools. A brief description of each facility follows.

Facility I: Production and Characterization of Proteins

A microbe's genome contains instructions for making the proteins that perform nearly all the functions of life, including those that can contribute to DOE missions such as energy generation, environmental cleanup, and carbon sequestration. If we can understand how proteins "do their work," we can use them to help solve these and many other problems.

Facility I will use highly automated processes to mass-produce and characterize proteins directly from genome information and affinity reagents ("tags") to identify, capture, and monitor the proteins.

Facility II: Whole Proteome Analysis

All the proteins encoded in the genome make up an organism's "proteome." The cell does not generate all these proteins at once but rather the set required at a particular time to produce those functions dictated by environmental cues and the organism's life strategy. To make use of any microbe's capabilities, we must understand the principles of these processes.

Facility II will characterize the expressed proteomes of diverse microbes under different environmental conditions as an essential step toward determining the functions and interactions of individual proteins and sets of proteins.

Facility III: Characterization and Imaging of Molecular Machines

Cells are biological "factories" that perform and integrate thousands of discrete and highly specialized processes through the coordinated use of molecular "machines" composed of assemblies of proteins and other molecules.

Facility III will isolate, identify, and characterize thousands of molecular machines from microbes and develop the ability to image component proteins within complexes and to validate the presence of the complexes within cells.

Facility IV: Analysis and Modeling of Cellular Systems

The final step in achieving a comprehensive understanding of living systems will require the ability to measure and predict dynamic events within individual cells.

Facility IV will combine advanced computational, analytical, and experimental capabilities for the integrated observation, measurement, and analysis of the spatial and temporal variations in the state of cellular systems—from individual microbial cells to complex communities and multicellular organisms.

Office of Science—At the Forefront of the Biological Revolution

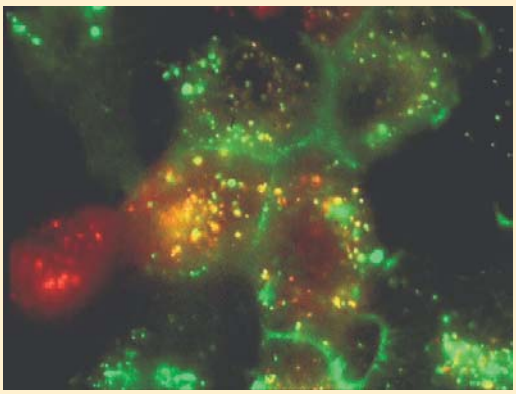
The knowledge and resources generated by scientists using the new facilities will provide GTL and the scientific community a powerful set of tools for conducting systems biology science, resulting in unique insights and opening new fields of scientific inquiry. These user facilities also will promote cross-disciplinary education of the first generation of scientists fully trained in systems biology and will attract faculty, postdocs, and students to GTL research.

Effective use of microbial and other biological systems and components will generate new biotechnological industries involving fuels, biochemical processing, nanomaterials, and broader environmental and biomedical applications.

The Office of Science has the capabilities and institutional traditions to bring the biological, physical, and computing sciences together at the scale and complexity required for success. Its academic affiliations, national laboratories, and other resources include major facilities for DNA sequencing and molecular-structure characterization, OASCR's high-performance computing resources, the expertise and infrastructure for technology development, and a tradition of productive multidisciplinary research essential for such an ambitious and complex program. In the effort to understand biological systems, these strong assets and the GTL program will complement and extend the capabilities and efforts of research supported by the National Institutes of Health, National Science Foundation, other agencies and institutions, and industry.

Lighting Up Proteins in Cells

Fluorescent labeling provides a way to study the functions of specific proteins in living cells. It allows direct observation in real time of the protein's location in time and space as well as the interactions of proteins with each other and with other cellular elements. The thousands of affinity reagents produced by automated methods in Facility I (see p. 2 and more detailed facility description beginning on p. 10) will provide scientists a means to study these components in living cells, both in GTL user facilities and in the scientists' own labs. For more information on this image, which shows the distribution and expression levels of two proteins in a grouping of cells, see p. 12.



Why Microbes?

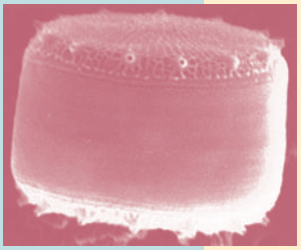
The ability of this planet to sustain life is largely dependent on microbes. They are the foundation of the biosphere, controlling earth's biogeochemical cycles and affecting the productivity of the soil, quality of water, and global climate. Microbial research is one of the most exciting frontiers in biology today, revealing the hidden architecture of life and the dynamic, life-sustaining processes on earth. Achieving an understanding of these secrets promises revolutionary solutions to many currently intractable energy and environmental problems.

Microbes must recognize available sources of energy to survive and thus have become masters at harvesting it in almost any form. Optimized to capture energy and materials for growth and cell maintenance, microbial cells can mitigate environmental threats such as toxins or extremes in pH, temperature, and salinity. Their life strategies enable microbes to carry out sophisticated biochemical functions for degrading wastes and organic matter, cycling nutrients, and, as part of the photosynthetic process, converting sunlight into energy and "fixing" (storing) CO_2 from the atmosphere.

The diversity and range of their environmental adaptations mean that microbes long ago "solved" many problems for which scientists are seeking solutions today. DOE's science missions require novel approaches for cost-efficient environmental cleanup, production of fuels (e.g., methane, ethanol, and hydrogen), and mitigation of global climate change. Microbes are capable of carrying out all these processes, but fully harnessing their natural capabilities first will require a complete understanding of their biological systems, the ultimate goal of Genomes to Life.

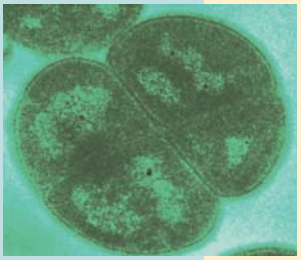
Microbes for DOE Missions

Carbon Sequestration



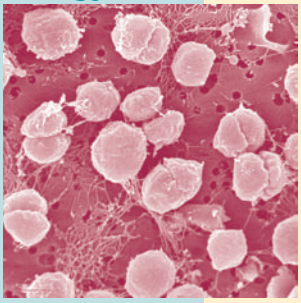
Thalassiosira pseudonana: Ocean diatom that is major participant in biological pumping of carbon to ocean depths.

Bioremediation



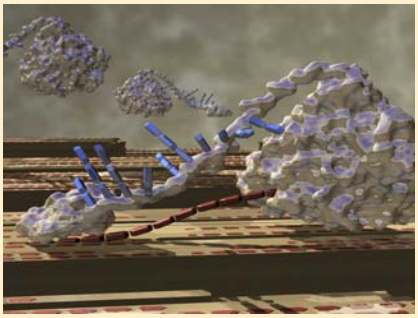
Deinococcus radiodurans: Survives extremely high levels of radiation and has high potential for radioactive waste cleanup.

Energy Production

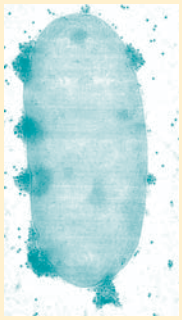


Methanococcus jannaschii: Produces methane; contains enzymes that withstand high temperatures and high pressure, possibly useful for industrial processes.

Cellulose Degradation



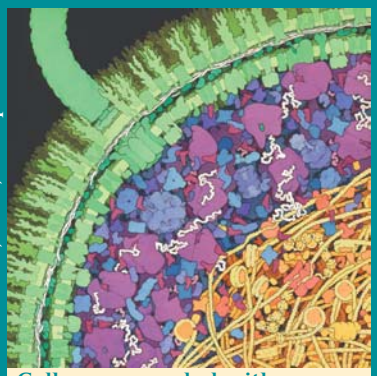
Cellulase molecular machine



Microbulbifer

Structures on the surface of *Microbulbifer* contain molecular machines (similar to cellulase, pictured above) that can break down cellulose in plant cell walls, an important step in converting cellulose to ethanol. See the description of Facility III: Characterization and Imaging of Molecular Machines to learn more about the science that can help enable this important energy application (p. 22).

Analyzing Microbes Requires Economies of Scale



D. Goodsell, 1999, with permission

Cells are crowded with many components, including proteins. A flagellum (green) protrudes from the surface. Note the molecular machine (also green) at the base that produces the flagellum's movement.

Genomes to Life will begin the analysis of a single biological system such as a microbe by using sequence databases to generate and study thousands of microbial proteins and the tags that allow them to be identified, captured, characterized, manipulated, and imaged in living systems. These resources will enable measurement of the spectrum of individual proteins, their assemblies ("molecular machines"), and associated metabolites that occur within cells and cellular communities under different experimental conditions. Analysis and modeling of cellular systems will combine knowledge of pathways, networks, and molecular machines to generate understanding of cellular and multicellular systems. High-performance computing will be used to assimilate and integrate the resulting data and information.

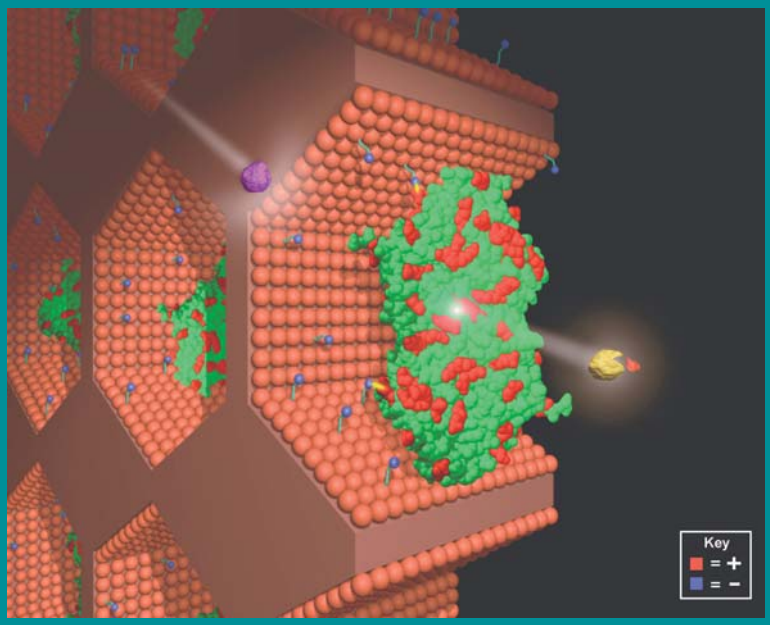
Large-Scale Facilities Spur Cost, Productivity Improvements

The dramatically increased productivity and reduced costs achieved in the Human Genome Project via high-throughput sequencing facilities provide the paradigm for the dedicated industrial-scale facilities envisioned for Genomes to Life.

Many elements of this analysis are practical only in large facilities. BER and OASCR propose the establishment of four high-throughput, integrated core user facilities to develop and implement these technologies (see text for details). These resources will create an integrated knowledge base for systems-biology science available to researchers worldwide. Applications of this level of knowledge will be far-reaching across the life sciences.

A Possible Application of Knowledge Gained in GTL Facilities

Harnessing Enzymes to Inactivate Contaminants, Generate Energy, Sequester Carbon



Miniaturization technologies can be combined with biological components such as enzymes to create novel systems that use an array of microbial processes but do not require living cells. In this figure, the enzyme organophosphorus hydrolase (OPH) is embedded in a synthetic nanomembrane (mesoporous silica) that enhances its activity and stability [*J. Am. Chem. Soc.* **124**, 11242–43 (2002)]. Applications such as this could enable development of efficient enzyme-based ways to produce energy, remove or inactivate contaminants, and sequester carbon to mitigate global climate change. It also could be highly useful in food processing, pharmaceuticals, separations, and the production of industrial chemicals. For more details on this illustration and large-scale protein production in GTL, see Facility I: Production and Characterization of Proteins, p. 10.

Genomes to Life Program Management

Pilots and Awards

Over the past several years, BER has funded pilot studies in technology development and systems biology. These projects have demonstrated mass spectrometric analysis of microbial proteomes, development of new imaging modalities, small-scale generation of microbial proteins, and development of computational tools for first-generation genome analysis and annotation. These pilots are producing data and experience to identify bottlenecks, areas for technology development, and scale of facilities needed.

In July 2002, DOE announced five major research awards for GTL systems biology to consortia involving many institutions (see below) and totaling $103 million over the next 5 years. These awards represent the culmination of nearly 3 years of planning by the DOE Office of Science and hundreds of scientists at universities, national laboratories, and industry. The microbes studied in the pilot projects as well as the 2002 awards have potential for bioremediating metals and radionuclides, degrading organic pollutants, producing hydrogen, sequestering carbon, and demonstrating importance in ocean carbon cycling. All have had their genetic sequences determined under DOE's Microbial Genome Program.

2002 Awards

- Oak Ridge National Laboratory and Pacific Northwest National Laboratory. "Genomes to Life Center for Molecular and Cellular Systems: A Research Program for Identification and Characterization of Protein Complexes"
- Lawrence Berkeley National Laboratory. "Rapid Deduction of Stress Response Pathways in Metal/Radionuclide–Reducing Bacteria"
- Sandia National Laboratories. "Carbon Sequestration in *Synechococcus*: From Molecular Machines to Hierarchical Modeling"
- University of Massachusetts, Amherst. "Analysis of the Genetic Potential and Gene Expression of Microbial Communities Involved in the In Situ Bioremediation of Uranium and Harvesting Electrical Energy from Organic Matter"
- Harvard Medical School. "Microbial Ecology, Proteogenomics, and Computational Optima"

Other Participating Institutions

Argonne National Laboratory

Brigham and Women's Hospital

Diversa Corporation

Los Alamos National Laboratory

Massachusetts General Hospital

Massachusetts Institute of Technology

National Center for Genome Resources

The Institute for Genomic Research

The Molecular Science Institute

University of California (Berkeley, San Diego, Santa Barbara)

University of Illinois

University of Michigan

University of Missouri

University of North Carolina

University of Tennessee (Knoxville, Memphis)

University of Utah

University of Washington

URL for Call for Proposals: www.er.doe.gov/production/grants/Fr03-05.html

GTL Web Site: DOEGenomesToLife.org

Fostering Research Community Participation

The Web site of the Genomes to Life program is designed to inform and foster participation in this exciting new undertaking by multidisciplinary investigators in the greater scientific community, science administrators, related policymakers, educators, and the general public. A suite of educational resources such as genome posters and handouts is accessible, and teachers can request multiple copies of materials. All GTL publications are posted, as are downloadable images, workshop reports, funding announcements, and abstracts of cutting-edge technologies.

Direct Web Access

- doegenomestolife.org/pubs.html
- doegenomestolife.org/gallery/images.html
- doegenomestolife.org/research/index.html
- www.ornl.gov/hgmis/education/education.html

Achieving a Molecular-Level Understanding of Life: A National Science Priority

OMB, DOE Budget Statements

The recent Office of Management and Budget's (OMB) Office of Science and Technology Policy memo, "FY 2004 Interagency Research and Development Priorities," declares that achieving a "molecular-level understanding of life processes" is a national science priority. The memo notes, "Sequence and structure data, coupled to modern computational power and to our ability to manipulate biological systems at the molecular level, will yield new experimental approaches that have the potential to unravel the complexity of life at the molecular, cellular, and organismal levels."

Solving DOE mission challenges requires the application of a systems approach to biology. DOE has the capabilities and institutional traditions to bring the biological, physical, and computing sciences together at the scale and complexity required for success in these efforts. The DOE FY 2003 budget request noted that "one of the most exciting areas of exploration is in the study of microbes—'bugs' that withstand extreme environments and may one day solve our energy-production problems and eat their way through our toughest environmental-cleanup areas."

AAM Recommends New Technologies

Specific recommendations from the American Academy of Microbiology's (AAM) 2001 colloquium report on "Microbial Ecology and Genomics: A Crossroads of Opportunity," are (1) "Develop new technologies, including methods for measuring the activity of microorganisms (at the level of populations and single cells); approaches to cultivating currently uncultivable species; and methods for rapid determination of key physiological traits and activities; and (2) "Establish mechanisms to encourage the necessary instrument development." A related recommendation is to (3) "Encourage instrumentation development through collaboration with device engineers, chemists, physicists, and computational scientists, since uncovering the diversity and activities of the microbial world is dependent on such advances."

Another recommended goal was to (4) "Develop technology and analysis capability to study microbial communities and symbioses holistically, measuring system-wide expression patterns (mRNA and protein) and activity measurements at the level of populations and single cells."

BERAC Approves GTL Facilities Strategy

Since the inception of the GTL program, the DOE Offices of Advanced Scientific Computing Research (OASCR) and Biological and Environmental Research (BER) have sponsored 15 workshops over the last 2 years to guide program implementation. Participants represented the breadth of the scientific community in industry, national laboratories, and academia. A strategic plan for developing facilities to serve the entire community was approved by the BER Advisory Committee (BERAC) in April 2002. The GTL roadmap is being updated in 2002.

Understanding Molecular Machines Requires Ultrascale Computing

Molecular machines—assemblies of proteins and other chemical components—carry out most of life processes. Computationally simulating molecular machine activity is a critical prelude to understanding and using microbial capabilities and requires levels of computing power beyond that available today. For more details on computational modeling of molecular machines, see sidebars, pp. 26 and 33.

User Facility Governance

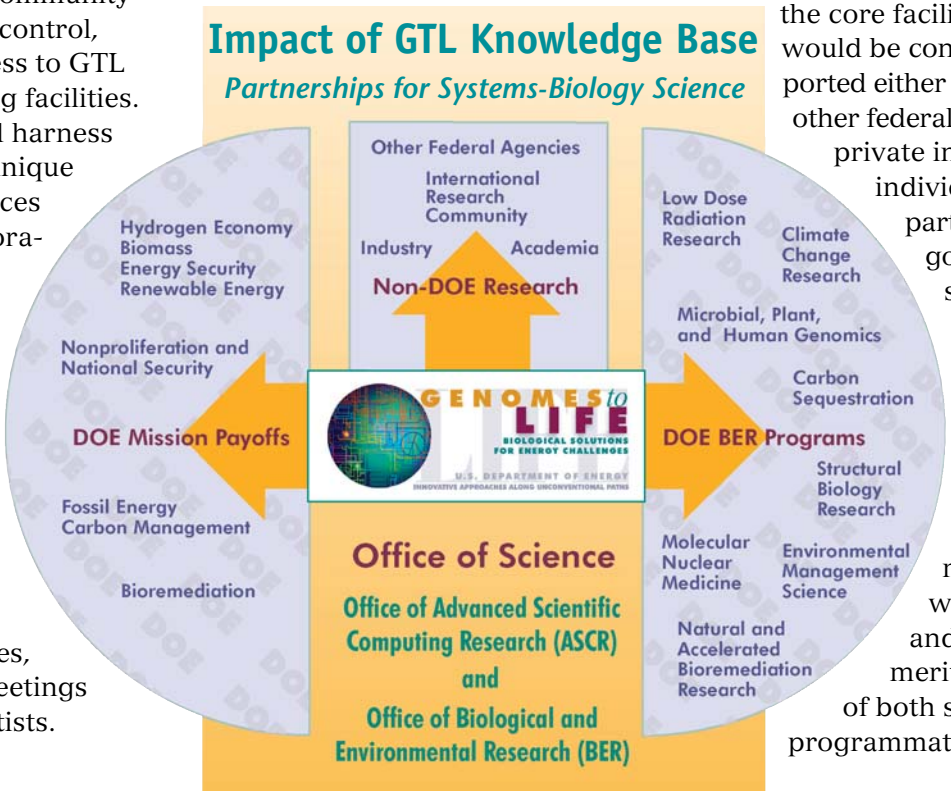
The goal of the GTL program is to explore the amazingly diverse characteristics and capabilities of microbes, gain insight into the life processes of cells and communities of cells, and understand functions that can be exploited for solutions to mission problems. A consensus is growing within the scientific community that achieving a higher level of biological understanding requires a new paradigm based on an integrated or systems approach—a synergistic application of experiment, theory, and modeling.

This new paradigm will require numerous and significantly more complex approaches and dedicated user facilities that provide the broader scientific community with technologies and computing and information infrastructure to gain the necessary innovation, efficiency, and efficacy. Biology will be democratized, and the most sophisticated and comprehensive capabilities, reagents, and data will be available to investigators lacking such integrated technology suites in their own laboratories or institutions. Moreover, new avenues of principal investigator–driven research will be enabled. Some of the GTL program's capabilities and facilities will transcend the program and will directly and indirectly benefit other public and private programs in systems biology.

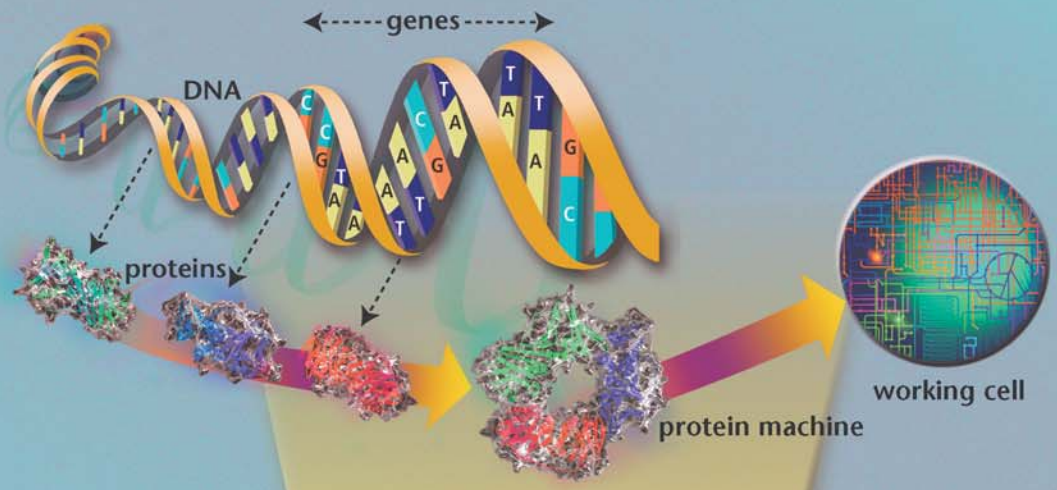
A sound governance and access model will ensure that the scientific community maintains positive control, influence, and access to GTL resources, including facilities. This enterprise will harness and integrate the unique powers and resources of the national laboratories, academia, and industry. DOE will seek advice from the scientific community through the usual mechanisms that could include the National Academy of Sciences, inter- and intra-agency advisory committees, workshops, and meetings of supported scientists.

Management and financing of programs have evolved over the years, particularly for facilities, and most user facilities are now managed with what is termed the "steward-partner model." This model was developed to ensure that user facilities provide the maximum scientific benefit to the broadest possible research community in the most cost-effective manner. It was implemented in the report, *Synchrotron Radiation for Macromolecular Crystallography*, Office of Science and Technology Policy (February 5, 1999). The model is described in some detail in the National Research Council report, *Cooperative Stewardship: Managing the Nation's Multidisciplinary User Facilities for Research with Synchrotron Radiation, Neutrons, and High Magnetic Fields* (National Academy Press, 1999). It also is followed in the recent report, *Office of Science and Technology Policy Interagency Working Group on Neutron Science: Report on the Status and Needs of Major Neutron Scattering Facilities and Instruments in the United States* (June 2002). (Reports: www.ostp.gov/Science/html/cassman_rpt.html, www.nap.edu/books/0309068312/html/index, and www.ostp.gov/html/NeutronIWGReport.pdf, respectively.)

Investment in GTL facilities would be a national commitment to enable relevant cutting-edge science that also supports DOE programs. In this model, DOE (the steward) would manage and fund the core facilities. Research would be conducted and supported either by the steward or by other federal agencies, industry, private institutions, or individual scientists (the partners). Principles governing the steward-partner model would be used to provide GTL facility resources to the scientific community. Good stewards of the investments and trust would determine use and access by objective merit-based peer review of both scientific quality and programmatic relevance.



From DNA Sequence to Living Systems:
Advanced Technology Facilities Critical to Genomes to Life



High-Throughput Technologies

Facility I: Production and Characterization of Proteins
- Use genome data to generate and characterize proteins, tags, and other resources needed to study microbes
- In vivo, in vitro, and chemical synthesis of proteins
- Multiple biophysical characterizations
- Computational tools for tracking and biophysical analysis
- Automated superannotation of genome data

Facility II: Whole-Proteome Analysis
- Measure proteome and metabolites for a cell or community systems under controlled conditions
- Gain functional insights by characterizing known and unknown dynamic processes to correlate proteins and machines that work together in a process
- Large-scale proteome analysis via mass-spectrometry
- High-capacity cultivation systems
- Optical analysis and cell sorting
- Computational tools for petabyte-scale data

Facility III: Characterization and Imaging of Molecular Machines
- Isolate the repertoire of molecular machines
- Characterize machines in terms of composition and molecular organization
- Biophysical and imaging characterization of molecular machines
- Biophysical and imaging tools
- Computational tools for molecular machine modeling and image analysis

Facility IV: Analysis and Modeling of Cellular Systems
- Couple knowledge of pathways, networks, and molecular machines to generate understanding of cellular and multicellular systems
- Measure structure and properties of a single cell in a population or community under controlled conditions
- Molecular machine imaging and characterization in living cells
- Culture and analysis of complex microbial communities
- Computational tools and algorithms for modeling cells

- **Comprehensive understanding of living systems**
- **Applications to DOE missions and across the life sciences**

Facility I: Production and Characterization of Proteins

A microbe's genome contains instructions for making the proteins that perform nearly all the functions of life, including those that can contribute to DOE missions such as energy generation, environmental cleanup, and carbon sequestration. If we can understand how these proteins "do their work," we can use them to help solve these and many other problems.

GTL Facility I will use highly automated processes to mass-produce and characterize proteins directly from genome information and affinity reagents ("tags") to identify, track, quantify, manipulate, capture, and monitor the proteins.

Strategic Intent

Virtually every cellular chemical reaction and physical function necessary for sustaining life is controlled and mediated by proteins generally organized into multiprotein complexes, or "molecular machines." High-throughput, automated protein and affinity-tag production and subsequent functional analysis of proteins and complexes will enable the study of chemical and physical interactions of proteins that underlie biology. A typical microbial genome has 2000 to 5000 genes that contain both the recipes for the production of thousands of proteins and the regulatory signals that control production. The cell does not generate all the proteins at once but rather the set required at a particular time for the functionality dictated by environmental cues and the organism's life strategy.

A systems-level understanding of cellular behavior will require experimental data for a significant portion of an organism's proteins. We must have the capability to produce essentially all the thousands of proteins encoded in each of many different genomes.

While selection of particular proteins may be simplified by comparative genomic analyses, production rates for proteins and their affinity tags will be many times higher than currently available. The associated characterization data for these reagents will better benefit the scientific community when it is generated under automated, high-throughput, and standardized conditions because it will be highly reliable and reproducible. Protein availability will enable production of affinity reagents to identify the components of molecular machines and to specifically capture, label, and track proteins in living systems. Data and reagents produced in Facility I will be invaluable resources for understanding molecular machines and cellular processes.

The overall objective of the facility is to produce milligram quantities of tens of thousands of full-length, functional proteins per year; generate multiple affinity tags for each protein; provide initial biophysical characterizations of each protein; and construct a comprehensive production and characterization database. Proteins, tags, protocols, and annotations produced in this facility will be motivated by the needs of the DOE GTL research program, DOE science missions, and the broader biological research community. Expression vectors, proteins, affinity reagents, and data produced will be foundational resources for all other GTL facilities and the entire scientific community.

Specialized, large-scale facilities are needed to achieve the necessary economy of scale and output of standardized characterization data associated with each protein and affinity reagent. This capability is an essential component of the nation's science infrastructure.

Project Purpose and Justification

Protein production is currently limited by economic and technological constraints. In the absence of significant automation, costs are prohibitive for genome-scale efforts ($10,000 to $25,000 per protein plus similar costs for tags). Current efforts costing hundreds of millions of dollars by individual-investigator approaches are focused on only a fraction of selected proteins. This protein subset and associated affinity tags typically are made independently in multiple laboratories at great expense. Unfortunately, characterization data about processes that worked and those that failed usually are not standardized, so user communities have difficulty obtaining reliable and comparable data to build models. Furthermore, many data associated with production, storage, and characterization are not preserved. The goal of this facility is to address these issues and, in particular, to develop comprehensive databases for processes that will aid automation and dramatically lower the costs of producing proteins and protein-affinity tags. Development of high-throughput production and characterization using robotics integrated with

computation will provide significant economies of scale and the requisite fidelity and quality. A centralized, computationally integrated facility also will maximize efficient use of resources.

Production Capability

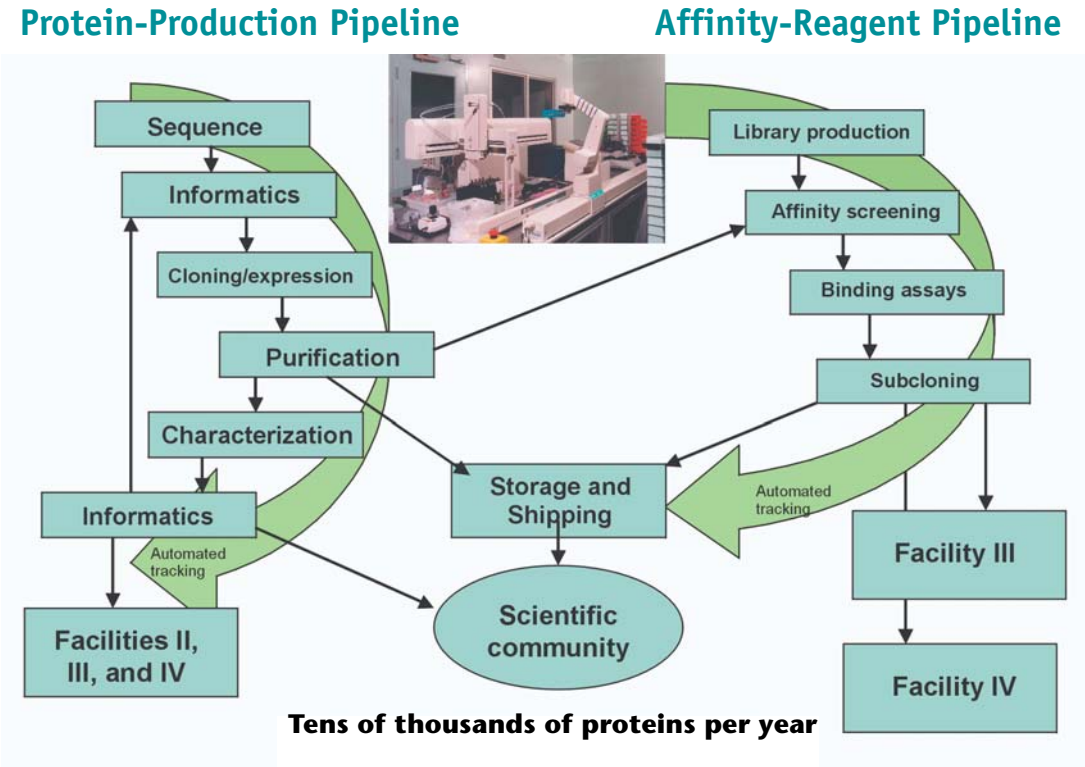
Scale

- 10,000 to 25,000 purified proteins per year
- Milligram quantities of each product
- Soluble, full length, and natively folded
- High rate of success (>95%) for production of proteins

Scope

- Proteins from genetic sequences of target organisms
- Protein variants
  - Isotopically labeled proteins
  - Post-translationally modified proteins
  - Proteins with unknown cofactors
  - Proteins incorporating nonstandard amino acids
  - Site-specific mutant arrays (high-throughput mutagenesis)
  - Fusion tag arrays
  - Affinity reagents (e.g., antibody domains) for every protein produced

Facility I: Production and Characterization of Proteins



GTL Facility I will use highly automated processes to mass-produce and characterize proteins directly from genome information and affinity reagents ("tags") to identify, track, quantify, manipulate, capture, and monitor the proteins. Protein production (left grouping) requires comparative, informatics-guided selection of appropriate targets, followed by cloning their genes, inducing expression in cell-free extracts or cells, and purification. Systematic biophysical characterizations will be completed on each protein. Once available, proteins will be used to generate specific affinity reagents (right). Current approaches rely on selecting affinity reagents from a diverse library and subsequent amplification. Expression clones, proteins, affinity reagents, and characterization data will be used by Facilities II, III, and IV to study molecular machines by mass spectrometry, imaging, and modeling approaches.

The currently funded NIH distributed structural genomics centers have had some success in producing proteins, and these results will be useful in developing Facility I. The focus of NIH efforts to exclude characterization beyond 3D structures means that many of the proteins selected for expression are either small or not full length and poorly represent the many types and classes of proteins. Key classes, comprising a third or more of the total, are significantly disordered in solution and thus are neither amenable to, nor meaningfully characterized by, conventional structural determinations. Yet, disorder in proteins is emerging as an increasingly important factor in determining function—particularly in the assembly of protein partners into molecular machines. This key process very often is mediated by disorder-to-order transitions at the binding interfaces. Facility I will provide general biophysical characterizations of full-length proteins that will, among other things, allow their general structure (whether ordered or disordered) to be defined. Integration of data obtained in Facility I, with appropriate informatics efforts, eventually will allow protein disorder to become a useful tool to predict binding partners and aspects of protein function.

The production of multiple high-affinity, high-specificity affinity-tag reagents for each protein presents its own enormous challenges. Several promising approaches to this problem are under development worldwide, although none have yet emerged as economical and reliable solutions to the high-throughput needs of GTL. Overcoming this obstacle is therefore a major target for GTL pilot studies and for this facility in particular. Computational tools will be employed to provide an initial understanding of the genes and protein complement of each microbe studied and to estimate protein function and organism capabilities based on the catalog of previously analyzed microbes. This analysis will identify novel proteins and those involved in previously characterized molecular machines. Prior knowledge and curated and archival data will be used to build predictive models of protein function and behavior. These data and models will be readily accessible to enable studies of proteins and protein complexes.

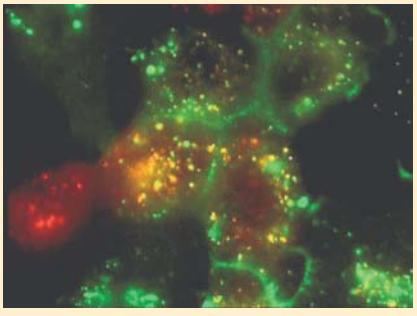
Proteins, tags, and data produced in Facility I are needed by Facilities II, III, and IV to capture the molecular machines for mass spectrometry (MS) analysis and to identify the machines' components. They also are needed for cellular-imaging studies, reconstitution of molecular machines, and verification of models. Characterization data will provide vital insights about which affinity tags are likely to disrupt or not disrupt cellular protein function.

Key Technologies Needed

- Capabilities for large-scale production and purification of milligram quantities of active, full-length proteins. This task includes difficult proteins to produce and characterize, such as those that are unstable or disordered, particularly in the absence of their binding partners.
- Reliable and high-throughput methods to successfully refold proteins.
- High-throughput methods to characterize proteins by multiple, independent biophysical assays under several standardized conditions.
- High-throughput methods to produce and characterize multiple affinity tags for each protein.
- Laboratory Information Management System (LIMS) for tracking and managing samples, tags, and production conditions, as well as sample export to the research community.
- Computational tools for efficiently collecting, analyzing, and interpreting the above-noted production and characterization data. These will

Lighting Up Proteins in Cells

Fluorescent labeling using tags such as those produced in Facility I provides a way to study the functions of specific proteins in living cells. It allows direct observation in real time of the protein's location in time and space as well as the interactions of proteins with each other and with other cellular elements. The thousands of proteins produced by automated methods in Facility I will provide scientists a means to study these components in living cells, both in GTL user facilities and in their own labs. This image shows the distribution and expression levels of the two proteins in a grouping of cells. The yellow emissions indicate simultaneous green and red emissions and thus reveal regions where both proteins are present (within the microscope's resolution of ~0.4 μm).



include tools for analyzing successful and unsuccessful expression; generating initial comparative genome analysis to determine genes, proteins, and regulatory elements; and identifying previously known protein associations with other proteins and ligands.

- Data-management systems for capturing knowledge about previously examined microbes; detailed protein, machine, and regulatory-element comparisons; and protein-production data and conditions.

Project Description

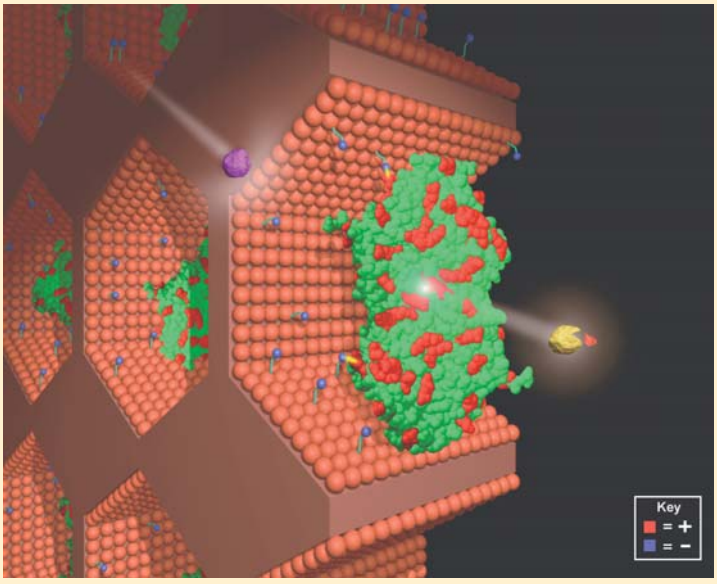
Facility I will revolutionize how proteins, affinity reagents, and associated characterization data become available to the scientific community. A sophisticated informatics capability tracking all aspects of production and characterization means that crucial data and reagents could be applied to a myriad of scientific problems. This computational infrastructure will enable use of the DNA sequence to predict the following for each protein: efficient and successful production methods, likely binding partners, and ultimately information about the functions of each gene. Achieving this goal will require experience and the data created from production and characterization of tens of thousands of proteins. Automation and computationally based insights are keys to achieving high throughput at steadily declining costs, just as they were in DNA sequencing. Proteins are more difficult to handle than DNA, so no single production method and characterization scheme will be applicable to all proteins. Thus, several methods will be developed simultaneously.

Whichever method is selected, nearly all protein production is based on transcription from DNA. This DNA is produced via cloning or possibly direct chemical synthesis of the gene encoding the desired protein. Facility I, as part of its function as a national resource, will develop a sequence-verified library of publicly available protein-coding microbial genes. This library would be available for translation into protein or usage in transformational studies by the other facilities or the larger scientific community.

A Possible Application of Proteins Produced in GTL Facilities

Harnessing Enzymes to Inactivate Contaminants and Generate Energy

Miniaturization technologies can be combined with biological components such as enzymes to create novel systems that use an array of microbial processes but do not require living cells. In this figure, the enzyme organophosphorus hydrolase (OPH) is embedded in a synthetic nanomembrane (mesoporous silica) that enhances its activity and stability [*J. Am. Chem. Soc.* **124, 11242–43 (2002)]. The pores (up to 30 nm) were functionalized with a few negative charges to aid in holding the positively charged OPH in place. The figure shows the enzymes, pores, functionalization (blue balls), and contact between negatively charged pores and some positively charged regions of the enzyme (glowing blue and red balls).**



A unique feature of OPH is its ability to inactivate an unusually broad range of chemicals (substrates), including several nerve gases (e.g., sarin) and pesticides. In addition to using individual enzymes, the technology also may be useful for immobilizing clusters of enzymes in particular metabolic pathways. Application such as this could enable development of efficient enzyme-based ways to produce energy, remove or inactivate contaminants, and sequester carbon to mitigate global climate change. It also could be highly useful in food processing, pharmaceuticals, separations, and the production of industrial chemicals.

High-throughput protein production requires multiple approaches based on cell-free and cellular methods. Direct chemical synthesis may someday represent a viable alternative, although refolding into active protein remains a major unsolved problem.

Cell-Free Systems

Cell-free expression systems, such as those based on wheat germ extracts or *Escherichia coli,* hold the greatest potential for full automation and hence lower costs and high throughput. Successful efforts in Japan using these extracts have yielded thousands of proteins per year. The ability to automate these systems and the potential to incorporate labeled or nonstandard amino acids warrant a substantial investment in these highly promising and flexible in vitro methods in Facility I.

Cell-Based Expression Systems

Large-scale cell-based expression systems have been used with some success worldwide in structural genomics centers and elsewhere. These approaches cannot, however, be as readily automated as cell-free systems. They also suffer from fairly low success rates. Partly for this reason, yeast and other eukaryotic expression systems have been developed and are typically resorted to for proteins that fail in *E. coli* –based systems.

Chemical Synthesis

Solid-state chemical synthesis is a possible approach for important proteins that fail in all DNA-based expression systems. Currently, this method can produce peptides up to 50 amino acids in length, but longer peptides are made at ever-diminishing efficiencies. Full-length proteins might be synthesized through chemical ligation of multiple peptides. This is, however, a costly procedure, and refolding into active protein remains a major unsolved problem. This technique has the advantage of producing milligrams of proteins labeled by incorporation of isotopes, chemical modifications, unnatural amino acids, or other chemical groups. Facility I will implement this approach in later years.

Protein Purification

Protein purification (*after expression*) presents a number of challenges, particularly in a high-throughput environment. In Facility I, substantial reliance will be placed on experience-based informatics methods to guide the purification strategy for each protein—with the expectation of achieving significant improvement as the database expands. Automated protocols aimed at eliminating centrifugation will be developed since this step accounts for the major bottleneck in current protein-production protocols.

Characterization of Proteins

A key and largely unique goal of Facility I is stabilization and extensive characterization of each produced protein under well-defined conditions. Given the investment in each expressed protein and its scientific value, subjecting each to a substantial suite of assays is planned: solubility measurements by light scattering; probing conformation and disorder by circular dichroism; HSQC nuclear magnetic resonance; partial proteolysis and isotope exchange coupled with MS; small-angle scattering, dye-binding and spectrofluorimetry; surface plasmon resonance; calorimetry; and protein chip–based binding measurements to extracts, proteins, ligands, and nucleic acids. This and the large number of assays contemplated and the high targeted rate of protein production imply the need for a careful and systematic development of high-throughput, robotic approaches. Facility I will be responsible for characterization data that is scientifically useful.

Affinity-Tag Production

Once proteins become readily available, they enable production of protein-specific affinity reagents. Several different and complementary approaches to generating affinity reagents are under development worldwide. These include phage and yeast display systems and aptamers. All have promising features, but none of these technologies has been developed sufficiently to satisfy the high-throughput requirements necessary for this facility. Further developmental areas include improved reagent stability and specificity; improved multiplex screening protocols; and rapid, high-throughput affinity-maturation techniques. The reagents also will be evaluated to determine where they bind to their protein targets and whether they disrupt the target's function, thereby dictating how different affinity reagents can be used. Development of modular affinity reagents also would be extremely useful; selected binding domains could be inserted into standardized structural modules to allow affinity reagents to be generated rapidly for different purposes such as protein isolation or live-cell imaging.

The most useful affinity reagents probably will be proteins themselves. They can be produced and characterized using technologies already developed for bacterial proteins. Because they will be standardized reagents, however, processes can be developed to allow for their rapid and large-scale production, enabling their distribution to scientists worldwide and greatly enhancing the scientific impact of reagents generated in the facility.

Computational Infrastructure

Central to this facility will be computational resources that provide an initial estimate of genes and proteins in a genome and identify proteins with known function or associations in previously characterized molecular machines. Computational tools will be utilized to estimate the capabilities of each new genome and help prioritize proteins for production. Systems are needed that will allow tracking of samples from the DNA constructs to the incorporation of data into gene-annotation databases. LIMS will be used to track samples (via bar coding) and incorporate all information relevant to sample history. A suite of computational tools for automated analysis and archiving of protein production and characterizaton data will be established to feed bioinformatics tools that will interpret these data.

Impacts on Science and DOE Missions

The postgenomic era of biology will focus on the genome and how it creates and utilizes proteins. Availability of thousands of microbial proteins and specific, high-affinity tags will substantially free the nation's scientists to apply their energies to understanding how microbial proteins function. Microbial capabilities then can be harnessed for exciting applications to DOE missions. Examples are hydrogen fuel cells based on hydrogenase enzymes supplied and maintained by clusters of associated molecular machines or environmental cleanup based on enzyme inactivation of toxic chemicals. This facility will provide unique resources and technologies important to systems biology.

Probabilities for Success

- DOE has the capabilities to establish a centralized facility that combines biological, physical, and computational sciences at the scale required for successful production and characterization of vital proteins and tags.
- DOE has supported many of the necessary components as pilot programs including the following:
  - Pilot facilities for protein production have been developed in conjunction with structural genomics research at several national laboratories including Argonne, Brookhaven, Los Alamos, Pacific Northwest, and Berkeley and are supported in partnership with the NIGMS Protein Structure Initiative. These facilities, which produce milligram quantities of hundreds of proteins per year, are developing the technologies needed to bring protein production to the next level of automation, completeness, and reproducibility.
  - The production of affinity reagents for proteins is being explored at several national laboratories, including Pacific Northwest, Argonne, and Los Alamos. These reagents cannot be designed but rather are selected from very large combinatorial libraries of reagents. The national laboratories are providing the necessary experience for developing novel libraries and automating their screening for the high-throughput production of affinity tags.
- Some aspects of protein production by cell-free systems can be modeled on facilities in Japan that are willing to assist DOE in establishing its significantly larger and more comprehensive Facility I.

Facility II: Whole Proteome Analysis

All the proteins encoded in the genome make up an organism's "proteome." The cell does not generate all these proteins at once but rather the set required at a particular time to produce the functionality dictated by environmental cues and the organism's life strategy. To make use of any microbe's capabilities, we must understand the principles of these processes.

GTL Facility II will characterize the expressed proteomes of diverse microbes under different environmental conditions as an essential step toward determining the functions and interactions of individual proteins and sets of proteins.

Strategic Intent

While the information content of the genome is relatively static, the processes by which families of proteins are produced and molecular machines are assembled for specific purposes are amazingly dynamic, intricate, and adaptive. Microbes deploy an interacting and changing panoply of proteins to carry out the myriad processes necessary for life. In addition to the networks and structures that do the cell's core work, numerous other machines and networks serve as sensors and regulators for the production and control of all the cell's elements. The principles for deployment of cellular capabilities can be deduced much more readily if a cell's protein makeup can be measured and correlated to external stimuli and to cellular responses. Thus, characterizing a microbe's expressed collection of proteins is an important first step toward deciphering the principles by which the genome regulates the assembly and functioning of molecular machines. However, a microbe typically expresses thousands of distinct proteins at a time, and the abundance of individual proteins may differ by a factor of a million. Technologies only recently have emerged that can successfully measure this breadth and dynamic range, and a substantial technical and computational infrastructure is required for their use (see "*Deinococcus*" sidebar, p. 18, on the FTICR Mass Spectrometer and Accurate Mass Tags results].

The Whole Proteome Analysis Facility (Facility II) will employ state-of-the-art techniques to generate data with high efficiency and validity; it will bring the same economy of scale to proteomics that centralized centers brought to gene sequencing. This facility will build on information obtained from the genome sequence to identify, in a snapshot fashion, the thousands of dynamically changing proteins expressed in a living microbe.

Project Purpose and Justification

The new era of systems biology requires the generation of massive amounts of whole-systems data collected under highly controlled and reproducible conditions from the point of bacterial cell growth to data archiving and dissemination. No facilities currently exist with the range and scale of capabilities and capacities required to collect these types of data. Further, centralizing the analysis of proteins within a specialized facility in a manner analogous to today's genome-sequencing centers would allow us to conduct these assays with higher efficiency, fidelity, and throughput than could be accomplished in the laboratories of individual investigators.

This facility will establish capabilities that will permit the annual measurement of the expression levels of proteins in bacteria grown under hundreds of different conditions. This will make possible the integrated study of thousands of proteins of numerous microbial species under a significant range of environmental changes. For the first time, we will be able to visualize a cell's systems-level response to these changes and identify the critical molecular changes that result from those conditions.

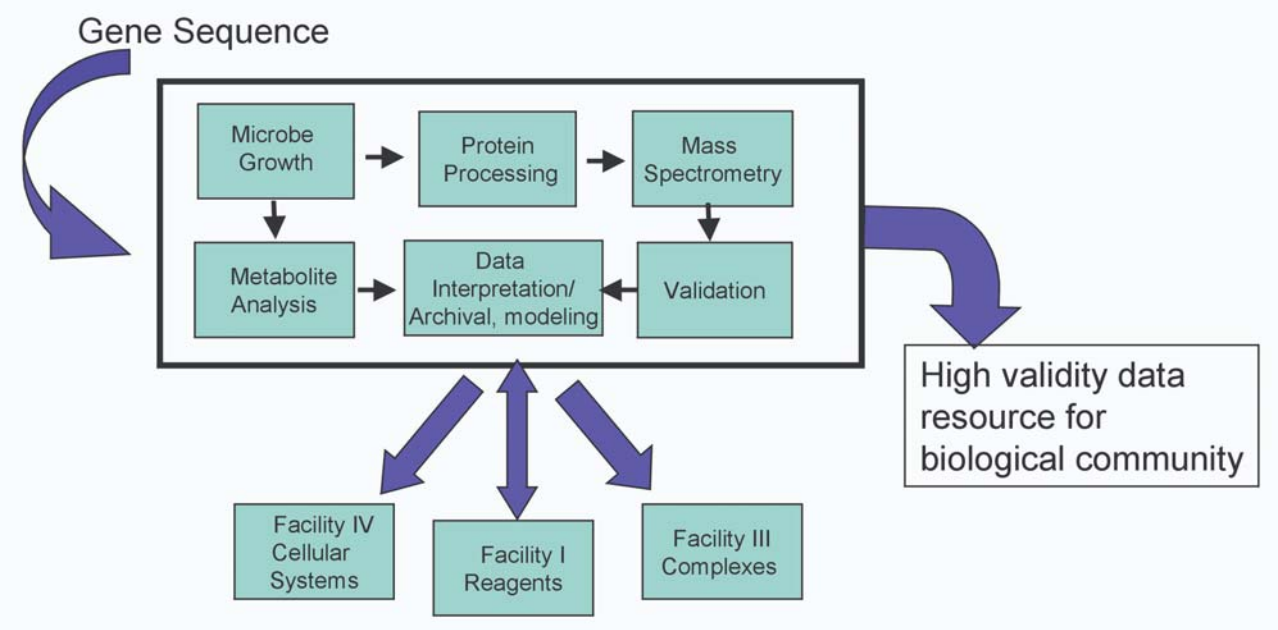
Many of the technologies needed for whole-proteome analysis have been successfully demonstrated in pilot projects funded by the BER program. Facility II will incorporate those technologies and others within suites of integrated analytical and computational tools. This facility will include capabilities to grow microorganisms under controlled conditions, isolate proteins from cells, and identify and quantify proteins using MS and other analytical techniques. Computational tools will be employed to interpret and archive data and to build predictive models of subsystems that control protein expression and affect cellular response to conditions. Computational modeling and simulation will be used interactively with experimental data collection to achieve a quantitative understanding of the components and parameters affecting expression. This comprehensive knowledge—cap-

tured in data, models, and simulation codes—will be disseminated to the greater biological community to enable studies of microbial systems biology.

Key Technologies Needed

- High-capacity cultivation systems, including controlled and automated chemostats for growing microbes under defined conditions in batch or continuous culture and capabilities to handle difficult-to-cultivate microbes.
- High-throughput techniques for preparing microbial samples before proteome analysis that incorporate integrated sample-processing systems, robotics, and automation.
- Capabilities for large-scale analysis of microbial proteomes, incorporating demonstrated MS-based methods to produce high-quality data. Analytical and computational methods for detecting and quantifying proteins modified by such methods as phosphorylation and methylation.
- High-sensitivity analytical tools for high-throughput analysis of metabolites, lipids, carbohydrates, and other cellular constituents.
- High-performance computational tools and codes for efficiently collecting, analyzing, and interpreting whole-proteome data. Tool capabilities, including data clustering, expression analysis, and genome annotation, would be closely linked to the advances in computing infrastructure being proposed by DOE.
- Computational tools for abstracting network and pathway information from expression data and genome annotation and for building mathematical models that represent subcellular systems responsible for protein expression and proteome state (including modified proteins) as a function of conditions. Simulation would be used to evaluate the state of knowledge contained in these models and validate the accuracy of experimental parameters.
- Data-management systems for archiving large amounts of data that may exceed petabytes and that can be accessed easily by a large community of users. Databases of expression measurements, metabolome measurement, and networks and pathway systems, models, and simulation codes

Facility II: Whole Proteome Analysis



GTL Facility II will characterize the expressed proteomes of diverse microbes under different environmental conditions as an essential step toward determining the functions and interactions of individual proteins and sets of proteins. Proteins identified in Facility II will help guide the production of reagents in Facility I. Reagents from Facility I also will be employed to assist in verifying the identities of proteins studied in Facility II. Proteomics data from Facility II will enable the identification of protein complexes in Facility III. Similarly, data from this facility will enable the study of proteins, metabolites, and other cellular constituents in the cellular systems studied in Facility IV.

would be developed.

- Research activities to develop new capabilities to improve the throughput, sensitivity, and information content of analytical tools and research in complementary computational methods to better interpret and visualize the results of MS measurements and other types of complex experimental data.

Project Description

This facility will consist of specially designed laboratories to house resources for cell growth; high-throughput sample preparation; state-of-the-art analytical instrumentation, including a suite of mass spectrometers; and computational infrastructure for sample management, data analysis (leveraging DOE's high-performance computing infrastructure), curation, archiving, and dissemination.

Proteome-Expression Systems

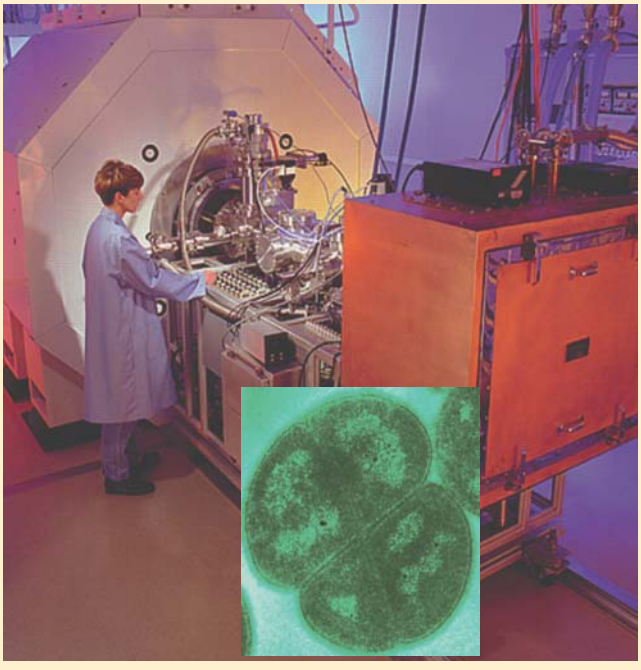
Cells will be grown under controlled conditions to produce proteins in sufficient quantities for analysis and would require fermentation technologies. The goal will be to rapidly assay proteomes from the target organism grown under a range of well-defined conditions. These multiple expressed proteomes will aid in rapid hyperannotation of new genomes by directly linking each genome to the expressed

A Possible Application of Knowledge Gained from GTL Facilities

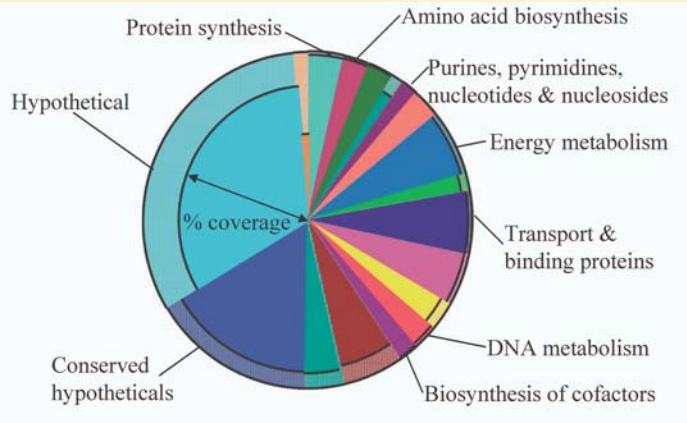
Deinococcus radiodurans: *Offering Promising Solution to Site Cleanup*

Dubbed the "world's toughest bacterium" by the *Guinness Book of World Records*, *D. radiodurans* (inset photo) can withstand extremely high levels of radiation and has excellent potential for use in stabilizing radioactive contaminants. Understanding the proteins involved in these capabilities may one day enable use of this microbe or its components in environmental cleanup. A BER pilot project to determine the proteome (the complement of proteins) of this microbe generated the most complete coverage obtained yet for any organism (*PNAS*, Aug. 20, 2002). Some 1900 proteins were identified, representing 83% of proteins predicted from the genome sequence (see pie chart below). The Office of Science's Microbial Genome Program provided the genomic-sequence information.

This unprecedented coverage was achieved using a new high-throughput mass spectrometer based on Fourier-transform ion cyclotron resonance (FTICR) developed at Pacific Northwest National Laboratory (photo). The system relies on a two-step process that first uses tandem mass spectrometry to identify biomarkers (accurate mass tags or AMTs) for each protein (see pie chart below). With this method, thousands of proteins were identified in a matter of hours. Identifying AMTs allows changes in the proteome to be monitored extremely efficiently.



Functional Classifications of Proteins



proteome under multiple growth and stress conditions. These whole-cell samples will be archived for subsequent analysis. Associated with these expression laboratories will be a range of analytical equipment, including chromatographs and mass spectrometers, to analyze other cellular components, such as metabolites. Significant investment will be also made in developing methods for working with microbes.

Protein-Sample Processing

Highly automated processes will be established to perform initial isolations of proteins from microbes, final sample preparation (e.g., desalting, buffer exchange, and sample concentration), and enzymatic digestion of samples as required for analysis. Sample-handling steps will be minimized using robotic and liquid-handling systems to eliminate many of the current bottlenecks in preparing samples for analysis.

Mass Spectrometry of Proteins

MS will be used to measure the molecular masses and quantify both the intact proteins and the peptides produced by enzymatic digestion of the proteins. Identification of the expressed proteins will require both moderate-resolution "workhorse" instruments, such as quadrupole ion traps, and high-performance mass spectrometers capable of high mass accuracy; the latter includes Fourier transform ion cyclotron resonance (FTICR) mass spectrometers or orthogonal injection time-of-flight mass spectrometers (called Q-TOF). These instruments will be interfaced with liquid chromatographs equipped with autosamplers to permit online separation of components prior to MS analysis. The data output of these instruments will require extensive dedicated computational resources for data collection, storage, interpretation, and analysis. [see "*Deinococcus*" sidebar on FTICR MS, p. 18]

Quality Assurance and Other Analytical Techniques

Quality assurance will be an important component within this facility. Validation of the identities and quantities of proteins present would be performed using complementary analytical and radiolabeling techniques and would require such equipment as high-performance liquid chromatography, spectrophotometers, gel electrophoresis, mass spectrometers, expression arrays, imaging tools, and computer workstations.

Computational Resources and Capabilities

Central to this facility will be an array of computational resources that will be employed to track samples and handle all aspects of data collection, storage, interpretation, and analysis. Databases and tools will be established for use by the biological community to access the data and models produced by the facility. LIMS will be used to track samples and incorporate all information relevant to sample history. A suite of computational tools for automated analysis and archiving of mass-spectral (protein-expression) data will be developed to feed bioinformatics tools that will interpret these data and identify proteins and their post-translational modifications. Such computational resources also will use data input from analyses established for quality assurance, including data from multiple MS runs, gel electrophoresis, and other assays.

This facility will develop and deploy large-scale data-analysis tools and infrastructure, tools for modeling and simulating protein expression based on collected MS data, petabyte-data management, and user interfaces for dissemination of data and proteome models to the community. For large-scale data analysis, modeling, and simulation, the facility will employ DOE's high-performance computing infrastructure. Additional data generated by other facilities also will be incorporated in these databases to enhance understanding of protein function in cells.

Optical Analysis and Cell Sorting

Biological systems are inherently inhomogeneous; measurements of the average proteome's expression profile for a collection of cells cannot be related with certainty to the protein-expression profile of any particular cell. This is especially true for proteins found in small amounts that may be expressed either at low levels in most cells or at high levels in only a small fraction of the cells. For these reasons, Facility II will use fluorescent probes developed in conjunction with Facility I to enable quantitative imaging of proteins within living cells. To achieve the most direct correspondence between imaging data and proteomic data, Facility II will conduct some of these measurements on the actual cultures used for the MS analyses. As a refinement, flow cytometric techniques will be used to separate various cell states to allow specific groups of cells to be studied from heterogeneous cultures. Other studies can be conducted using imaging capabilities in the other GTL facilities and at universities and national laboratories.

Technology Development

An important part of this facility is the development of new biological, analytical, and computational tools to improve the sample throughput and information content required for the GTL program. Although current state-of-the-art techniques allow the analysis of many proteins within a cell, new approaches will be required to permit efficient analysis of the full range of proteins without special handling. These approaches would include methods for isolating and analyzing membrane proteins, improving methods for quantification, and developing technologies for analyzing proteins from a single cell. Few computational tools are available to analyze these types of data and translate them into functional information and robust models of cellular subsystems. Many research activities need to be conducted in close association with this facility to meet the needs of the GTL program; numerous innovations, however, will come from individual investigators across the scientific community.

Impacts on Science and DOE Missions

Many microbial processes may be useful for DOE missions such as energy production, carbon sequestration, and environmental cleanup. Relating the microbe's genome to the specific functions conducted by its proteins may make possible the activation or deactivation of particular processes. This and other capabilities will enable us to optimize microbial systems or learn how to construct nonliving systems that mimic processes found in microbes for specific applications—a precursor to harnessing such processes to meet the needs of DOE's missions.

Understanding changes induced in a microbe's protein-expression profile by different environmental conditions will serve as a basis for identifying the function of individual proteins. This will provide the first step toward understanding the function of the complex network of processes conducted by a microbe.

Data from Facility II will be used to develop models for predicting microbial responses to different environments and for using these capabilities in practical applications that meet DOE mission needs. Further, if we can understand the processes in relatively simple microbial cells and extend this understanding to communities of microbes, then we can apply this knowledge to higher organisms.

Probabilities for Success

- Experience in the Human Genome Project taught us that biological analyses, like DNA sequencing, can be standardized, automated, and made high throughput and cost-effective. Similarly, large-scale protein analysis can provide high-quality data to the general scientific community.
- Research under way in the Genomes to Life program, industry, and across the federal government will identify many individual technology components and methods needed to make large-scale proteome analysis practicable.
- A BER-supported pilot research project already has characterized more than 80% of the computationally predicted proteome of one microbe under a variety of experimental conditions. Similar work is under way on *Shewanella*, *Rhodopseudomonas*, and *Prochlorococcus*.

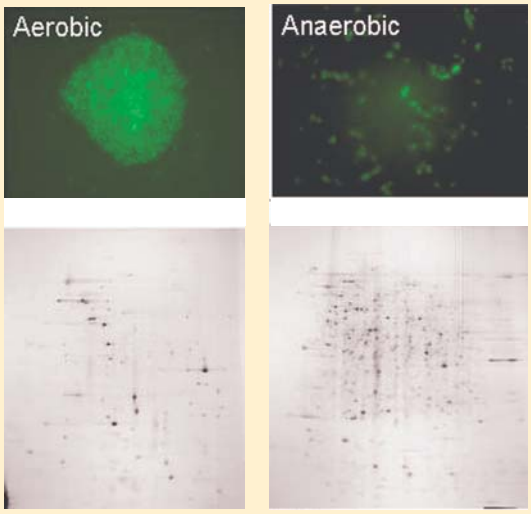
As this facility progresses, multidisciplinary teams of physical, biological, and computational scientists will continuously expand its capabilities to improve the throughput and information content of analytical data and provide new computational models for predicting protein expression in microbial cells.

A Possible Application of Knowledge Gained from GTL Facilities

Shewanella oneidensis: *Offering New Strategies for Groundwater Bioremediation*

Confocal microscopy images

2D protein gels



Proteome of aerobic and anaerobic *Shewanella* cultures

The ability of *S. oneidensis* to precipitate radionuclides (e.g., uranium and technetium) and metals (e.g., chromate), rendering them immobile in sediments, offers exciting new opportunities for developing new groundwater bioremediation strategies to clean up DOE legacy wastes. A GTL pilot project involving a consortium of scientists is now exploring how the microbe senses and responds to its environment. The team is cultivating the microbe under variable but carefully controlled conditions and measuring global (population) changes in gene expression and the resulting proteome. This work is enabled by the availability of the microbe's genome sequence, determined in the Office of Science's Microbial Genome Program.

Pictured is the growth of cells (top row) and protein complement (bottom row) present in the microbe under different environmental conditions. The pattern of cellular growth appears very different under aerobic (minibiofilm) and anaerobic (individual cells) conditions. These differences are reflected in the proteins revealed by 2D gels of the two cell cultures. An ability to quantitatively and comprehensively identify the cellular proteins will be critical for determining how *S. oneidensis* responds at the whole-system level.

Facility III: Characterization and Imaging of Molecular Machines

Cells are biological "factories" that perform and integrate thousands of discrete and highly specialized processes through the coordinated use of molecular "machines" composed of assemblies of proteins and other molecules.

Facility III will isolate, identify, and characterize from microbes thousands of molecular machines and develop the ability to image component proteins within complexes and to validate the presence of the complexes within cells.

Strategic Intent

Proteins seldom act in isolation; instead, they combine to form multiprotein complexes that function as "molecular machines." The genome and its associated sensing and regulatory networks control the creation and operation of the numerous molecular machines that make up a cell. These molecular machines are in turn organized into numerous tightly packed and highly interconnected physical structures. The dynamic interactions of proteins comprising many hundreds of these molecular machines are coordinated in time and space and are responsible for signaling, transport, motility, cell division, and virtually all other cell activities. Before we can progress toward a systematic understanding of cell function, we must discover which molecular machines can be produced by the cell under specific conditions and how they are positioned in the cell's structural architecture.

This facility's core role is to build on the data and reagents provided by Facility I and patterns in protein expression from Facility II to develop a detailed descriptive understanding of how proteins are organized into "molecular machines" and to locate the machines in the cell. The data from Facility III—together with dynamic snapshots of the complete, condition-dependent, expressed proteome to be generated by Facility II—will constitute the foundation on which Facility IV will develop a predictive, systems-level understanding of microbial cells and communities.

As important as protein complexes are in cellular function, our current knowledge of molecular machines is quite limited, partly because proteins most often have been studied individually and in isolation. Inherently difficult to study, many complexes are short-lived, unstable, or variable in their composition. In addition to identifying complexes, characterizing interactions among components will be critical to understanding their function. A first step toward this goal can be taken by determining the proteins comprising each complex and how the proteins interact to affect the machine's function.

Facility III will identify molecular machines and their components, characterize the interactions of the protein components of the complexes, and validate the occurence of these within the cell context. To accomplish this for the comprehensive set of molecular machines within the cell, the facility must develop automated analytical techniques for purifying, identifying, and characterizing multiprotein complexes—particularly, advanced imaging techniques. Integration, organization, and analysis of the data generated in this facility will require further development of principles, theory, and new computational tools for modeling and simulation of the structure and function of the complexes. Moreover, this facility will use the vast wealth of data on individual proteins being produced by structural genomics programs in other agencies including NIH and NSF.

Project Purpose and Justification

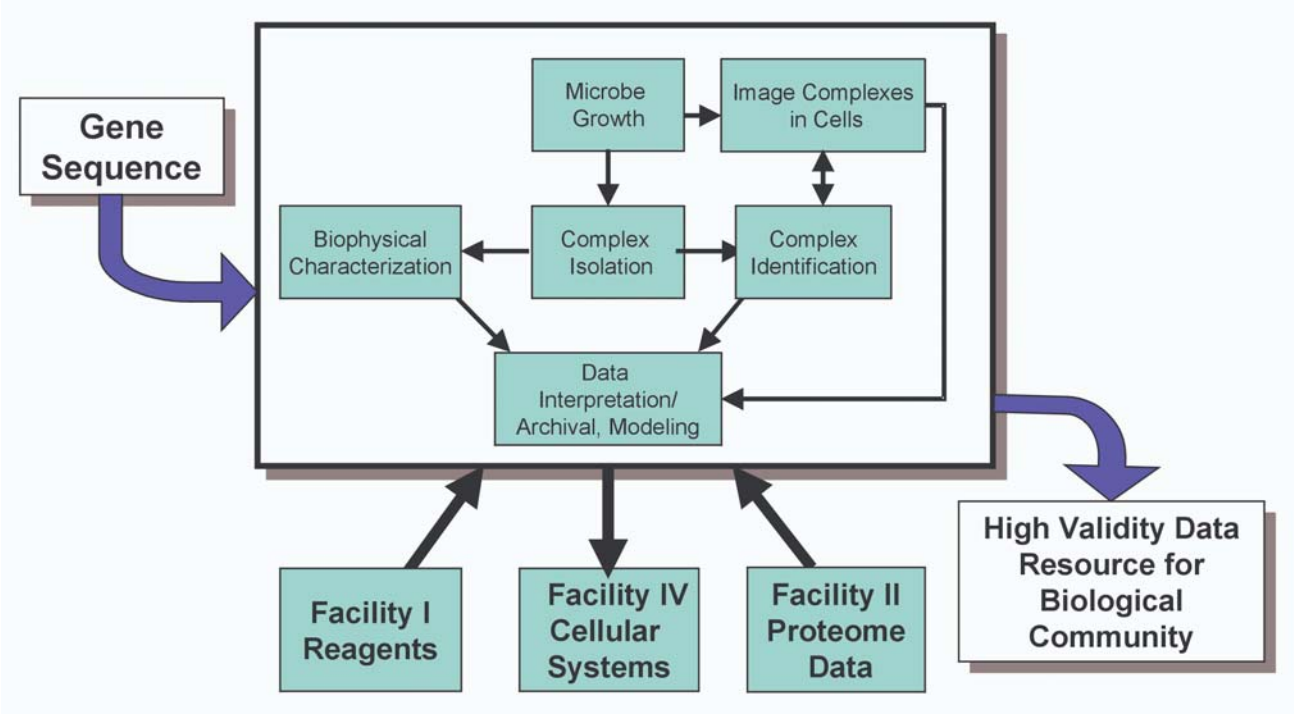
Molecular machines are highly dynamic, changing in composition, modification state, and subcellular location to carry out the vital functions of a cell. Responsible for a hierachy of molecular processes within and between cells, they dictate how a cell or organism interacts with its environment. A first step in determining how the network of cellular molecular processes works on a whole-systems basis is to completely understand individual molecular machines, how each machine is assembled in 3D, and how it is positioned in the cell with respect to other components of cellular architecture. Moreover, understanding how molecular machines operate at the molecular level will unlock the capability to control useful biochemical processes in a microbe and apply them to DOE mission needs.

This facility will focus on the isolation, identification, and characterization of thousands of molecular machines per year, providing revolutionary new information and capabilities to the biological community. Specifically, the facility will

- Discover and define the repertoire of protein complexes in a comprehensive manner; in this, it will rely critically on the information and reagents produced by Facilities I and II.
- Characterize the complexes as to their basic biophysical properties and also their inter-protein geometries.
- Develop molecular-level models to help interpret experimental data on protein interactions and modes of action of multiprotein machines.
- Develop theoretical principles for systematically describing the structure, aspects of function, assembly, and disassembly of multiprotein complexes.
- Validate the occurrence of complexes within the cells.

To reach the goal of the Genomes to Life program, vast numbers of data sets must be acquired from organisms maintained under a variety of well-defined conditions, then analyzed and made available to the biological community. Centralizing these analyses within a specialized facility analogous to today's genome-sequencing centers would allow assays to be conducted with higher efficiency, fidelity, and cost-effectiveness than could be accomplished in the laboratories of individual investigators.

Facility III: Characterization and Imaging of Molecular Machines



Facility III will isolate, identify, and characterize from microbes thousands of molecular machines and develop the ability to image component proteins within complexes and to validate the presence of the complexes within cells. Facility III requires the proteins and reagents from Facility I for the isolation, quantitation, and imaging of protein complexes. Facility II will provide baseline proteomics data for optimizing conditions for production of machines to be analyzed in Facility III. The detailed structural and biophysical characterization of the complete repertoire of a cell's molecular machines, provided by Facility III, is critical to understanding and modeling cellular systems in Facility IV.

Project Description

Facility III will house state-of-the-art analytical instrumentation for the identification and characterization of molecular machines. The instrumentation would include electron, optical, and force microscopes; mass spectrometers; and other analytical tools. Laboratories also will be required for microbial cell growth, molecular biology, high throughput, automated sample preparation, gene expression, mass spectrometry–based protein complex analysis, imaging of protein complexes, biophysical characterization, and quality assurance. Integrated with these facilities will be computing resources for sample tracking; data acquisition, storage, and dissemination; algorithm development; and modeling.

For multiprotein machines with structurally characterized components, high-performance computing will play a very significant role in constructing structural models of machines and performing molecular dynamics simulations of protein-protein interactions in molecular machines. The next generation of massively parallel processors in the 40- to 100-teraflop range will allow simulations of sufficient size and fidelity to make important contributions to explaining the mechanisms of machine construction and function.

In summary, Facility III will

- Isolate complexes from cells using high-throughput techniques.
- Identify molecular components of the complexes.
- Determine basic biophysical properties of the complexes.
- Interpret, annotate, and archive data for use by the greater biological community.
- Develop models for the assembly and activity of complexes and verify this information with experimental data.

Molecular Machine Isolation

Perhaps the most challenging task in the analysis of molecular machines is the isolation of these complexes from the cell. Protein complexes often are held together by only weak interactions, making them fragile and difficult to isolate for analysis. Many such complexes are present only briefly or in very low amounts—sometimes just a few per cell. No adequate techniques now exist for the robust, high-throughput isolation of protein complexes. The development and automation of such techniques is therefore an essential early goal of a current GTL pilot project for this facility. Data and reagents to be produced by Facility I will be central to isolating the multiprotein complexes. In particular, Facility I reagents such as antibodies or clones, intended to produce "tagged" proteins in Facility III, will be used to purify complexes from cells by "pull-down" experiments. These methods, however, must be highly automated to meet the ultimate goals of comprehensively identifying the multiprotein machines in a cell. Automated techniques will be established for final purification (i.e., desalting, buffer exchange, and sample concentration), stabilization, storage, and proteolytic digestion of samples as required for analysis. Novel techniques for analyzing protein complexes in single microbes also will be developed.

An important component of this facility is a highly integrated LIMS that will track samples from cell cultivation through data archiving.

Mass Spectrometry

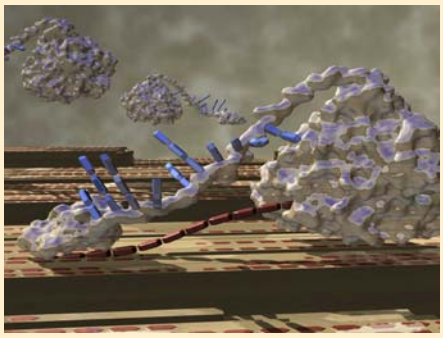
MS techniques provide the primary approach for analyzing the components of a molecular machine. This application, however, presents significant challenges. Spectrometers are required that add high-throughput operation to the current state-of-the-art standard in sensitivity, dynamic range, and resolving power. Mass spectra obtained from molecular complexes will produce data that can be evaluated on the basis of genomic-sequence data from the organism producing the complex and from proteomics data generated in Facility II. Typically, tens of thousands of spectra will be necessary to identify all protein components of a single multiprotein machine. Extensive computational analyses using genomic and proteomics data will be necessary to interpret these vast amounts of mass spectra from the complexes.

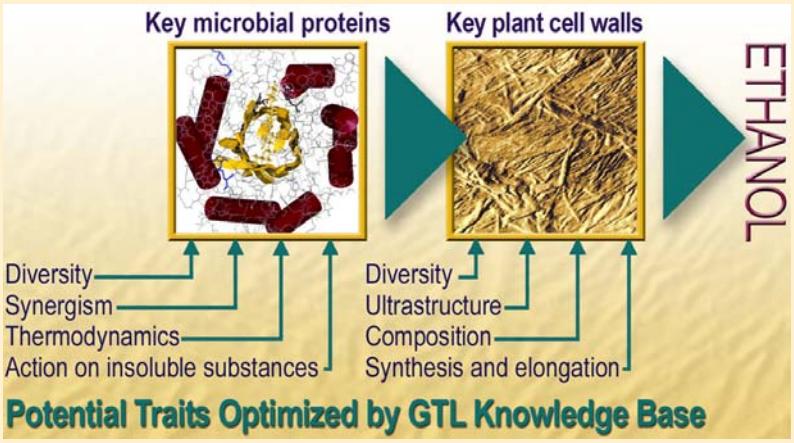
Biophysical Characterization

Generating isolated molecular complexes offers a unique but extremely challenging opportunity to characterize the complex with a host of biophysical techniques toward the ultimate goal of fully understanding the multiprotein machine's activity and mechanisms. Initially, a suite of well-established techniques will be employed to characterize the basic biophysical properties of an isolated complex. These techniques include surface plasmon resonance to investigate intermolecular interactions between the complex and weakly bound ligands; small-angle X-ray and neutron scattering to characterize overall structure and compactness of the complex; and wide-angle

A Possible Application of Knowledge Gained in GTL Facilities

Clean, Sustainable Energy





Enhanced plant qualities and microbial bioprocesses can be used to generate clean, sustainable energy. Microbial protein "machines" can break down the cellulose in plant cell walls for fermentation to ethanol. Today, the process is too inefficient for commercial production. Fundamental knowledge of gene regulation and protein machines gained in GTL can be applied to develop highly efficient methods to support large-scale ethanol production and displace a significant amount of fossil fuel use.

X-ray scattering to characterize secondary structure and folding. New technologies just being developed—such as single-molecule spectroscopy—are expected to allow more complete mechanistic understanding of multiprotein machines. Obtaining systematic experimental information about the dynamic behavior of the complexes (including their assembly and disassembly), combined with ongoing improvements in computational hardware and modeling methods, will allow accurate simulations of the activities of multiprotein machines at the heart of cellular function.

Imaging

Two different technologies will use (1) very high resolution to derive detailed 3D information about the complexes and (2) cell imaging to localize these complexes in individual cells.

Detailed 3D information about the structural organization of isolated molecular complexes will require many imaging technologies, including cryoelectron microscopy and a diversity of scanning and force microscopies. These tools will allow us to formulate the 3D structure of the complexes and provide hints as to how proteins interact.

An important application of imaging tools will be to verify the formation of complexes identified by MS and to map their location in the cell. Affinity reagents acquired from Facility I can be used to tag specific components of the complex and identify the location of complexes within the cell as well as the dynamics of assembly and disassembly. Electron tomography, X-ray microscopy, and live-cell imging using various light microscopy techniques will yield this data and will validate (in the cell) the protein interactions inferred from MS, biophysical, and 3D structure-imaging results. This information will provide additional insight into understanding the function of protein machines and will furnish valuable data for system-wide studies to be conducted in Facility IV.

Imaging methodologies produce vast amounts of raw data that must be analyzed computationally to produce interpretable images. Computer systems capable of acquiring and processing huge data streams must be assembled, and new algorithms must be developed to analyze and integrate data from multiple imaging modalities. Finally, strategies must be designed for archiving and distributing image data to the biological community.

Technology Development

As outlined previously, available methods for the isolation and MS-based analysis of protein complexes are not immediately adaptable to very high throughput operations. Thus, a key component of this facility will be the development of more robust and automated biological, analytical, and computational tools to improve sample throughput and information capture from these techniques.

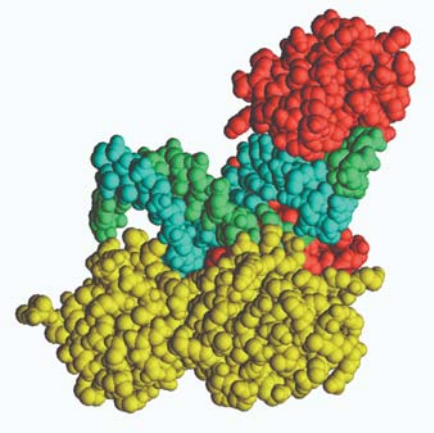
Impacts on Science and DOE Missions

Facility III will enable a fundamental understanding of the repertoire and properties of molecular machines present in cells. This is a prerequisite to determining how a microbe's network of molecular machines controls biochemical processes and therefore microbial life. It is the foundation on which we will achieve a fundamental scientific understanding of microbial life as well as the practical mastery needed to address DOE's mission goals in environmental remediation, energy production, and carbon cycling.

Probabilities for Success

- Analytical technologies have long been a strength of DOE national laboratories, and new technologies to analyze simple protein complexes are already in development in pilot projects funded by BER.
- DOE laboratories have a long history of excellence in high-performance computation and predictive simulations that can be applied to the study of molecular machines.
- Ongoing developments across many agencies and disciplines in robotics, automation, and data-management techniques will provide improved technologies. These technologies will be used to analyze the thousands of molecular machines in a microbe and identify biochemical pathways and gene regulatory networks that confer specialized capabilities possessed by microbes studied in GTL.
- Cryoelectron microscopy is a well-proven technique for analyzing the structure of molecular machines. Its development over the past 25 years has involved significant efforts at several national laboratories, including Brookhaven and Berkeley. High-end computing will enable integration of thousands of molecular images into progressively higher-resolution images of these complexes.
- New strategies using MS are being developed for isolating, stabilizing, and analyzing molecular complexes in one of GTL's initial projects. Oak Ridge National Laboratory is employing automated, high-throughput technologies and computational tools to capture these complexes for study. Partners are Pacific Northwest, Sandia, and Argonne national laboratories; University of Utah, and University of North Carolina.

Understanding Molecular Machines Requires Ultrascale Computing

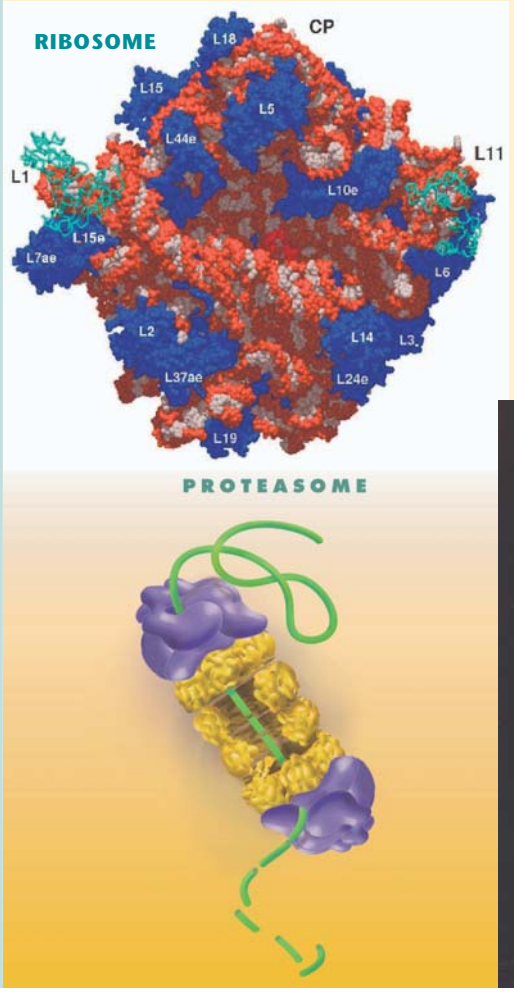


Computationally simulating molecular machine activity—a critical prelude to understanding and using microbial capabilities—requires levels of computing power far beyond that available today. Simulating just the key steps occurring during the activity of the DNA-binding complex in *Pyrococcus woesei* (image at left) would require 40 teraflops of computing muscle—beyond the limit of what scientists can do now. The machine is made up of transcription-initiation factor II B (yellow), the TATA-box binding protein (red), and DNA (green and blue). Obtaining a complete simulation of the full activities of this or other complex molecular machines would require a capability of more than 100 and perhaps as many as 1000 teraflops, which are not available today (see sidebar, p. 33), but are on the planning horizon for OASCR.

Tiny Molecular Machines Outstrip the Best Synthetic Chemists

Molecular machines—assemblages of proteins and other chemical components—underlie the dynamic life of a cell. They work together in functional networks to carry out most life processes, including executing metabolic functions, mediating information flow within and among cells, and building cellular structures.

The ribosome (top image, below) is the molecular machine responsible for protein synthesis in cells. This remarkable molecular machine translates the sequence of bases on a messenger RNA molecule (a transcript of the DNA sequence) into a parallel sequence of amino acids that make up a protein chain. The ribosome carries out all the chemical reactions needed to perform these tasks.



On an atomic scale, the ribosome is huge. Made up of more than 50 proteins and 3 or 4 strands of RNA, it contains over 100,000 atoms. Understanding its function—determined by the ribosomal structure—was one of the hottest challenges in biology from the 1970s until a few years ago, when the first high-resolution structures began providing some insights into how the ribosome conducts its activities. This work required herculean efforts by hundreds of scientists in numerous laboratories. Data collected at the synchrotrons at Brookhaven, Berkeley, and Argonne national laboratories were used to calculate the high-resolution 3D structure of the ribosome's large and small subunits. The availability of genome data and current mass spectrometry technologies would now enable this result to be achieved at a fraction of the time and cost.

In addition to making proteins, cells also must have a way to destroy or recycle them. Sometimes this is important for maintaining a balance of proteins and amino acids as cells respond to external stimuli. At other times, cells need to degrade misfolded proteins to reuse their amino acids and also to keep the defective proteins from interfering with cellular activities. A cell must have a carefully regulated way of deciding which proteins to degrade, so it makes sense that all cells have special molecular machines—proteasomes—to degrade proteins. Made up of roughly 30 proteins, this protein complex also is very large. The structure of the central portion has been solved at high resolution by protein crystallography, and the whole structure has been imaged with cryoelectron microscopy (photo at left).

Facility IV: Analysis and Modeling of Cellular Systems

The final step in achieving a comprehensive understanding of living systems will require the ability to measure and predict dynamic events within individual cells.

Facility IV will combine advanced computational, analytical, and experimental capabilities for the integrated observation, measurement, and analysis of the spatial and temporal variations in the state of cellular systems—from individual microbial cells to complex communities and multicellular organisms.

Strategic Intent

The ultimate empirical test of our understanding of genomes will reside in our ability to model and simulate entire living systems, from individual microbial cells to complex microbial communities and, eventually, multicellular organisms. Understanding how individual cells within populations and in multi-organism communities interact and function as a unit to carry out complex processes is key to unlocking their vast potential for applications of importance to DOE.

The ability of our planet to sustain all life is completely dependent upon microbes. They are the foundation of the biosphere, controlling biogeochemical cycles and affecting the productivity of our soils, the quality of our freshwater supplies, and local and global climates. Microbes carry out sophisticated biochemical functions to degrade wastes and organic matter, cycle nutrients, and, as part of the photosynthetic process, to convert sunlight into energy and "fix" CO_2 from the atmosphere.

In nature, microbes often live in communities containing many different species. Yet, biologists traditionally have studied microbes one species at a time in nutrient-rich media, conditions typically very different from the organism's native habitat. To complicate matters further, microbes can exhibit substantial cell-to-cell phenotypic variation even in pure culture, necessitating the measurement of cell properties and functions at the level of the individual cell. Biologists are now constrained by the ability to make measurements on individual living cells and in microbial assemblies have acknowledged that instrument development and new facilities are critical needs in microbial science.

To make the "final ascent" to a systems-level understanding of life, new and innovative capabilities are needed for the comprehensive characterization of dynamic cellular systems at the level of the individual cell and in the context of their environment.

To this end, Facility IV will combine advanced computational, analytical, and experimental capabilities for the integrated observation, measurement, and analysis of the spatial and temporal variations in the state of microbial cells. To achieve a systems-level understanding, simulation and modeling must be tightly coupled with experiment to define and analyze the complex regulatory and metabolic networks in microbial cells and communities. Facility IV, building on knowledge and materials from Facilities I through III, will provide new analytical and computational tools and infrastructure that will enable unprecedented insights into the cellular state of microbes in populations and, ultimately, communities and multicellular organisms.

Facility IV will utilize advanced instrumentation and will be data and compute intensive, providing linked data sets on the dynamic function and behavior of single-cell and multiorganism assemblages and the capabilities for developing and evaluating them.

In summary, Facility IV will

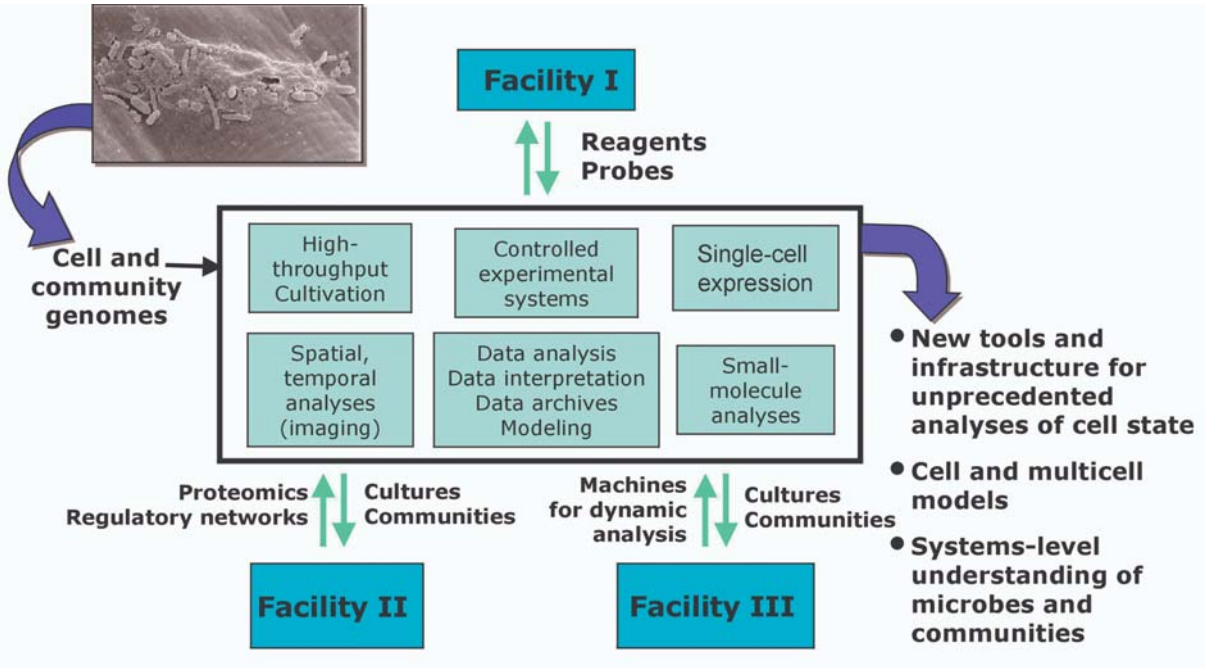
- Develop highly controlled systems for growing and maintaining microbial populations and communities.
- Model, simulate, and predict the responses of microbes to each other and their environment by developing high-performance computational algorithms and infrastructure.
- Examine the dynamics of molecular machines in living cells.
- Measure gene expression and track the locations and interactions of proteins within living cells.
- Understand how individual cells function and interact within complex microbial communities to carry out complex processes.
- Integrate experiment, analysis, and theory in a recursive fashion to reveal intra- and intercellular networks.
- Enable development of microbial technologies and applications to solve DOE mission-specific problems in energy, environment, and health.

Project Purpose and Justification

High-throughput gene sequencing is being applied to determine the collective "genome" of microbial communities as a first step in understanding microbial community structural processes. Genome sequence alone, however, is insufficient to understand the functionality of microbes. A key to solving the formidable problem of understanding microbial genome function is through intensive experiment, analysis, and theory in a coupled and recursive manner. In a recent American Academy of Microbiology colloquium titled "Microbial Ecology and Genomics: A Crossroads of Opportunity," instrumentation development and new facilities were identified as critical needs in microbial science (see "AAM Recommends New Technologies," p. 7).

Facility IV will enable integration of the information and material outputs of the other GTL facilities to bring genomes to life. The facility will provide the tools, capabilities, and infrastructure needed to predict the functional behavior of whole cells and communities as integrated, dynamic systems. Facility IV will emphasize the concurrent and dynamic measurement of cellular proteins, molecular machines, intracellular metabolites, regulatory molecules, and gene transcripts to establish the state of cells within populations and communities and as a function of changes in physicochemical biological conditions.

Facility IV: Analysis and Modeling of Cellular Systems



Facility IV will combine advanced computational, analytical, and experimental capabilities for the integrated observation, measurement, and analysis of the spatial and temporal variations in the state of cellular systems—from individual microbial cells to complex communities and multicellular organisms. Facility IV will require the integrated information, materials, and capabilities provided by Facilities I-III as well as whole-genome sequence from the Joint Genome Institute. Facility I will provide critical affinity reagents and tags for single-cell measurements, while Facility II will furnish data on regulatory networks and high-throughput quantitative proteome and metabolome measurements. Facility III will produce information on key molecular machines as a basis for investigating their dynamics and functions in living cells. Facility IV, in turn, will provide insights into the function of molecular machines by defining their locations and dynamic behaviors within living cells. Facility IV also will be a source of microbial cultures and the knowledge about how to cultivate them.

Key Technologies Needed

- Cultivation and maintenance of microbes and microbial communities under controlled conditions, including the ability to interrogate the function of individual microbial cells in the context of a characterized physicochemical environment.
- Novel high-throughput cultivation approaches combined with single-cell analysis techniques, such as emerging microfluidic "lab-on-a-chip" devices, to grow and study currently uncultivable members of microbial communities.
- New multimodal capabilities for dynamic imaging of molecular machines and metabolites in individual cells and cell assemblies coupled with advances in computational resources for data acquisition, storage, and analysis to interpret and visualize the vast quantity of information obtained.
- Probes for the in situ measurement of extracellular metabolites and for defining the physicochemical environment in near real time.
- Analytical instrumentation and techniques for identifying and characterizing spatial and temporal variations in metabolites and signaling molecules, within and surrounding living microbial cells and cellular assemblies.
- Computational tools for efficiently collecting, analyzing, and integrating large data sets to elucidate gene function and to model and simulate regulatory and metabolic networks.
- New theory, algorithms, and implementation on high-performance computer architectures, such as those provided by the Ultrascale Simulation effort, to model and simulate cellular systems.
- Web- and grid-based technologies to enable a broad range of biological scientists to access the large data sets and computational resources needed for discovery-based biology.

Project Description

Facility IV will provide multiple capabilities to the scientific community. It will serve as a focal point for teams of scientists from academia, industry, and government to conduct integrated experiments on microbial processes and systems of interest and allow them to develop the detailed high-quality data sets under strictly controlled and characterized conditions required for elucidating regulatory and metabolic networks. Technologies for analyses at the single-cell level will require unique high-end instrumentation coupled to controlled cultivation systems with capabilities and tools for intensive data collection, integration, and analysis. These capabilities will be used in part or in their entirety by individuals or teams needing comprehensive analyses, at the single-cell or community level, of cellular systems of interest. This facility also will serve as fertile ground for training the next generation of scientists in systems biology and will attract students and faculty to work with multidisciplinary teams to reveal the functions of the microbial and microbial-community genomes.

Computational tools will be used to analyze large data sets and develop detailed models of cellular systems based upon these measurements. Computational methods ultimately will be used for large-scale simulations with these models to provide the ability to predict the behavior of cellular systems. These new capabilities will enable biologists to exploit emerging high-end computational tools for data analysis and the development of reliable predictive models. Facility IV will allow the fulfillment of GTL program goals for a systems-level understanding of microbes and microbial communities relevant to DOE missions in carbon cycling, metal and radionuclide bioremediation, and biomass conversion.

Impacts on Science and DOE Missions

Facility IV will be the capstone facility needed to provide the knowledge synthesis critical for bringing genomes to life. Understanding how individual subsystems of cells and individual cells within microbial communities function in concert to sense, respond to, and modify their environment represents a grand challenge for biology that must be addressed before scientists can successfully predict the behavior of microbes and take advantage of their functions.

In comparison to GTL Facilities I –III, which will provide new high-throughput production and analysis capabilities focused on defining and understanding the parts of microbial systems, Facility IV will focus on the dynamic systems-level study of living cells. It therefore will be highly data intensive, providing extensive linked data sets on the dynamic behavior of microbial cells and communities.

It also will be compute intensive, providing unprecedented data analysis, modeling, and simulation. These systems-level data sets will be made available to

the scientific community and will be invaluable for identifying regulatory and metabolic networks in microbial systems and for advancing the annotation of microbial and community genome sequence. In addition, Facility IV will provide models and technologies for developing and evaluating such models and will provide user-facility type resources in the form of infrastructure needed to undertake such tasks.

Probabilities for Success

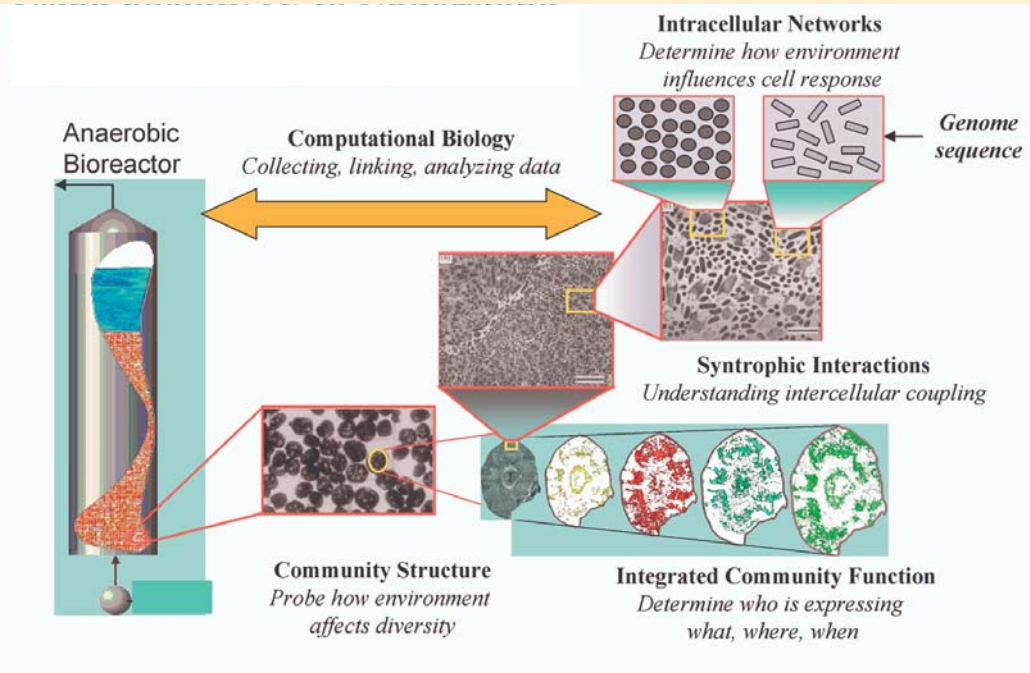
As in the genome projects, DOE can draw on multidisciplinary teams of biological, physical, computational, and other scientists and engineers from the national laboratories, academia, and industry to develop and deploy the resources for systematic functional genomic investigations of microbes and

A Possible Application of Knowledge Gained in GTL Facilities

Understanding Microbial Community Function is Essential for Using Bioreactors to Convert Waste to Clean Energy

Anaerobic bioreactors are used to treat wastewater and convert waste to "biogas" (mainly methane) that can be used as a clean energy source. Certain types of digesters contain self-aggregating biogranules composed of multiple spatially distinct microbial communities. The most widely used anaerobic treatment technology in the world, bioreactors have been the subject of extensive engineering studies. The biology, however, has largely been treated as a black box. The structure and function of microbial communities residing within the granules are believed to be important in maintaining an overall balance between production and consumption of metabolites and intermediates.

In Facility IV, the microbes, their molecular machines, and the complex interaction networks within and between cells in these communities would be revealed in detail. For example, the structure of biogranule communities, as well as integrated community function, would be established by determining which members are expressing particular genes or proteins at a given location and time. The integrated capabilities provided in this facility would allow scientists to "drill down" into these biogranule communities to probe intercellular coupling such as the transfer of metabolites that occurs between syntrophic microbial partners. Finally, at the level of the individual cell, scientists could determine how environment influences cell state and the dynamics of molecular machines within individual cells. A robust computational environment will be critical to Facility IV in collecting, linking, and analyzing experimental data and modeling and simulating complex systems such as the biogranule communities. The resulting information would have profound implications for controlling the efficiency and stability of methane-producing reactors and for greatly improving their design and operation.



microbial communities. DOE has an extensive and successful history of developing and applying new technologies to complex problems in the physical and chemical sciences. A tremendous opportunity now exists for applying these same talents to provide technological solutions to biology's most complex problems.

GTL Facility IV, more than the other three facilities, must take on and overcome major challenges associated with the lack of available technologies and instruments for measuring the dynamic state of living microbial cells. Facility IV will benefit from current and future R&D and pilot projects that will develop new technologies and instrumentation in a phased manner. These projects will provide new state-of-the-art capabilities by the time this facility comes on line. Facility IV also will include extensive capabilities for instrument and technology development that will be an essential part of this resource so that it can continue to measure the activities and characteristics of cellular systems at the single-cell level. In addition, it will involve development of new computational approaches for data storage, analysis, and use in complex models. For example, GTL-supported R&D and pilot projects under way include:

- Development of capillary analysis technologies to permit the monitoring of changes in protein expression in single cells using fluorescence, pushing the resolution by ten times over existing technology, is under way at the University of Washington. The goal is to build a "better microscope" for tracking gene expression in single cells following environmental changes.
- Electron tomography approaches are being developed at Lawrence Berkeley National Laboratory to image the inside of a microbial cell by freezing intact microbial cells in a way that preserves one layer of liquid water molecules above their membranes (permitting survival and viability). Electron microscopy images and computer reconstruction can then be used to derive 3D images of internal cell constituents.
- A pilot Microbial Cell Dynamics Laboratory is under development at Pacific Northwest National Laboratory to provide flexible experimental systems to control and manipulate microbial growth conditions and to make multiplexed measurements of cellular activities and responses to changes in environmental conditions.

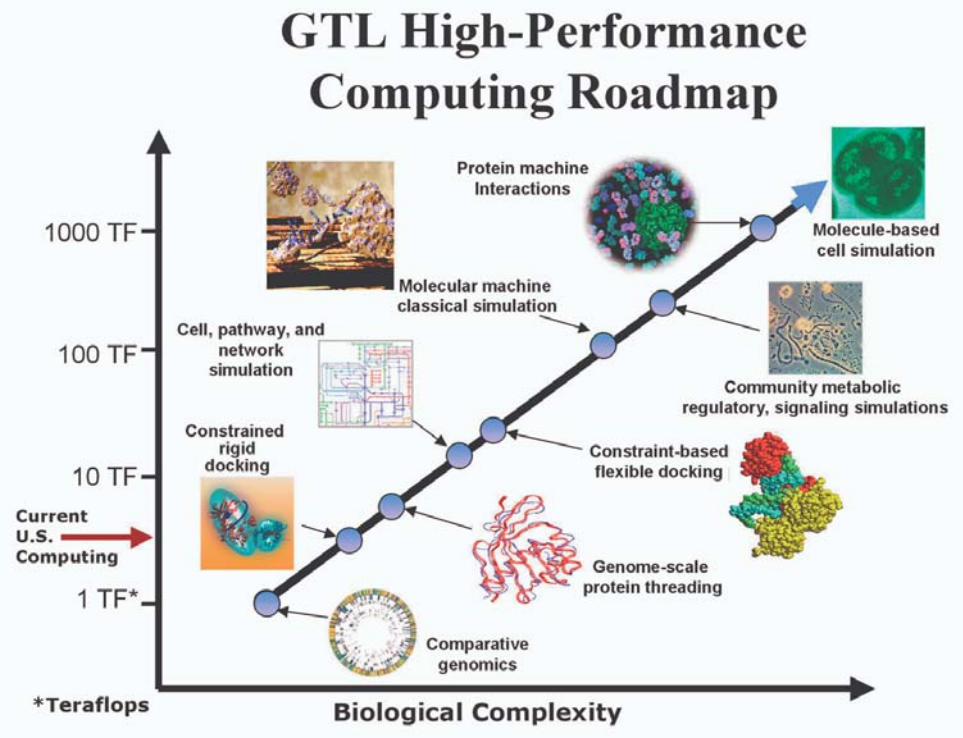
Cross-Cutting Computing Infrastructure for GTL Facilities and Projects

Strategic Intent

Biology is poised on the threshold of a great transformation that will create a new life-science approach in which large data sets and advanced computing will be combined into predictive simulations to guide and interpret experimental studies. This new biology will allow a level of understanding that will enable biological solutions to many of the world's pressing challenges including energy security, control of atmospheric carbon, and environmental cleanup.

Such progress depends on the emergence of a new mathematical, quantitative, predictive, and ultimately systems-level paradigm for the life sciences. This new paradigm is one in which biologists represent their most fundamental knowledge of complex biological systems as mathematically based computer models. These models will be used to capture and represent data, predict behavior, and generate hypotheses that can be tested by gathering more data (see "Biology Paradigm," p. 34). Biologists need to be provided with the means to move data and knowledge back and

Important breakthroughs in GTL modeling can be made using the next generation of high-performance computing platforms with 40- to 100-TF capability, including docking and simulation of protein-protein interactions using classical energy functions. More complex models and simulations of machines and cellular systems would benefit greatly from fundamental research in mathematics and from related algorithms that could dramatically improve calculation efficiencies compared to current estimates. Constraints (bounds and guides) provided by data such as observed machine geometries will make calculations of more complex systems tractable.



The so-called First Principles Molecular Dynamics (FPMD) methods, the current state-of-the-art in biophysical modeling, simulate the motions of atoms in biochemical systems using a quantum mechanical description of atomic interactions. Highly optimized for DOE's current teraflop-speed computers, FPMD applications are capable of simulating up to a few hundred atoms for a few picoseconds. These methods, however, also constitute a nearly exact simulation of nature, and, even within these computational limitations, FPMD is becoming an important tool for studying fundamental biochemical processes. The results for small biochemical systems currently being simulated on teraflop-scale computers provide tantalizing glimpses of the value of longer-time and larger-system simulations that will be made possible with faster computers.

forth between experiment and computing on an everyday basis, making large-scale computing an integral part of their daily lives. This will be necessary to study even the simplest microbes at a level of detail sufficient to predict their behavior. To ask next-generation questions and do next-generation experiments, computing must guide the questions and interpretation at every step.

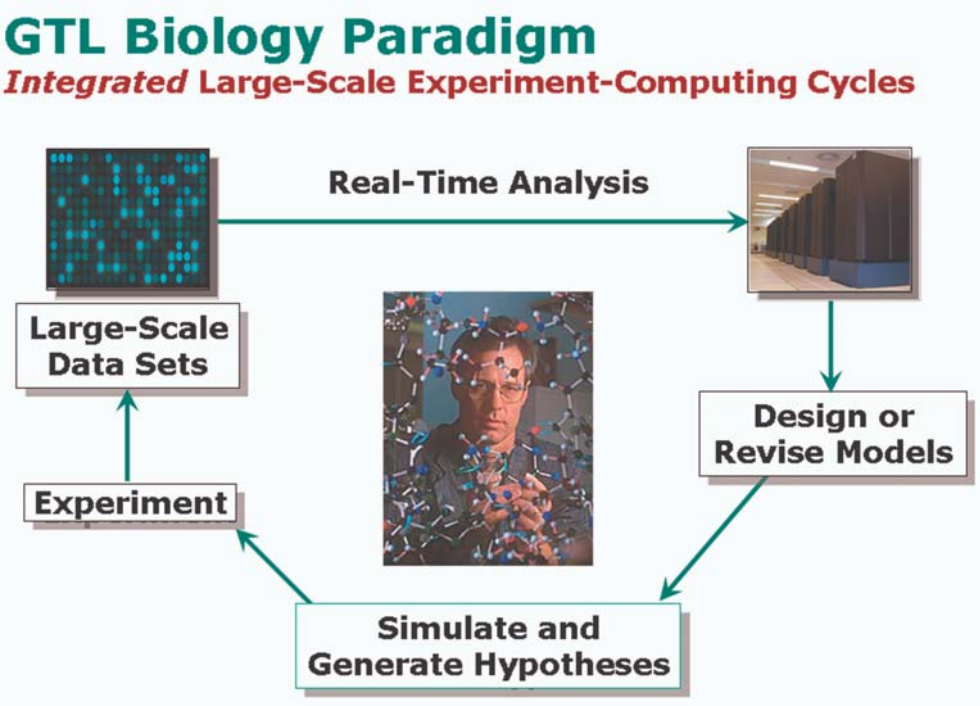
Adopting this new paradigm to meet the requirements of GTL facilities is the result of the following specific drivers.

- **Data Analysis.** Production facilities in Genomes to Life will generate vast amounts of diverse and complex data that must be analyzed, integrated, and interpreted. The complexity of data-generation modalities is much higher than in the genome era, and the amounts of data will be much larger than for sequencing the human genome. Achieving the necessary data-analysis throughput will require significant advances in software and hardware infrastructure.
- **Complexity.** The complexity of systems in even the simplest microbes makes manual analysis and annotation methods simply inadequate. This mind-boggling complexity needs to be captured in the computer and denote biology in mathematical ways that parameterize system complexity. The need to capture, represent, and model complex biology via computer language is fundamental to progress in Genomes to Life facilities and projects.
- **Prediction, Quantitation, and Simulation.** Quantitating, predicting, and simulating behavior are necessary to understand biological systems and develop hypotheses for further testing. Genomes to Life goals will require simulation of heterogeneous biological systems over long time scales and include molecular complexes, pathways, networks, and, eventually, communities of organisms. Quantitation is needed to test our understanding and representation of systems.
- **Principles and Concepts.** To understand the astounding complexity of biological systems, general systems principles need to be extracted and developed from systems data, modeling, and simulation. Mathematical representations of complex biological systems are fundamental to the conceptual breakthroughs anticipated in Genomes to Life.

Purpose and Justification

Computing capabilities for analyzing, modeling, and simulating the dynamic behavior of complex biological systems are an essential complement to GTL experimental data-collection activities that provide fundamental observations and data. In addition to providing a totally new capability for understanding the basis of life at its most fundamental level, the ability to compute and simulate large macromolecular machines and systems is central to many DOE missions, including bioremediation, climate, and energy security.

In bioremediation and climate, interactions of microbes with the environment and with each other also involve large complexes of macromolecules and metabolic systems that act on pollutants and sense environmental and community conditions. Data analysis followed by modeling and simulation can provide fundamental understanding as well as quantitative parameters important for models of larger earth systems such as climate.



In energy security, the potential for biological energy production is very great. Our understanding and ability to engineer complex systems that synthesize useful compounds would be enhanced by a fundamental grasp of the proteins and metabolic processes involved and the ability to predict the behavior and dynamics of these molecules and systems.

Software, Data, Biocomputing Centers

The Genomes to Life facilities plan and workshop reports place considerable emphasis on developing methods for a large community of biologists to analyze petascale biological data sets and develop models and simulations related to complex biological phenomena. They stress the need for centrally managed approaches to software and hardware infrastructure to accomplish these objectives. A centralized approach to coordination and planning in computing will guide data management, large-scale development of analysis tools, implementation, and support of analysis on specialized hardware environments, including massively parallel computers and distributed grid systems.

Components of the software-development and data-management infrastructure, such as GTL databases, will be designed around and co-located with major facilities that generate biological data. Experience in many large-scale genome-era projects, such as the Joint Genome Institute, has shown that physical co-location is immensely useful and absolutely necessary for effective communication and requirement definition among biologists and computational scientists. While the overall goal is to create a seamless and effectively centralized capability to deal with data, software development, and high-performance computing, key development teams will be based primarily at GTL experimental data facilities and coordinate their activities across the GTL enterprise.

The need is significant for large-scale compute tools and resources that will become a shared resource for the GTL community. These "centers" will, in many cases, be distributed within several GTL facilities or other sites. Three different types of components are required: (1) coordinated centers for applications software and tools development, (2) data-management and -integration resource centers, and (3) biocomputing centers that support community access, analysis, modeling, and simulation using a specialized computing hardware.

Several key points about this strategy are the following: (1) Domain experts who team to develop applications do not have to be co-located at sites where substantial computer hardware exists. (2) Centralizing management of applications ensures that they are supported, shared, and documented. (3) Analysis, modeling, and simulation applications can find cycles from multiple sites and are not as vulnerable to individual machine failures. (4) Although many users will be experts and understand the machine requirements of their codes, most biological users will not. To facilitate wide usage of GTL infrastructure in computing, very simple user environments must be created that "know" where an application should run (on what type of platform) and where to get the necessary data without the user having to specify these details. (5) By sharing compute hardware resources across the GTL enterprise and among GTL facilities, administrative processes can more effectively use a variety of machine environments, from large clusters to massively parallel processing (called MPP) machines, as the demand for processing dictates. Applications can be matched to the most appropriate environments.

Centers for the Development of Analysis, Modeling, and Simulation Tools

GTL data generation and computing advances will provide scientists with access to comprehensive information and the tools to incorporate it into models to probe the processes and phenomena of living systems, test hypotheses and ideas, and inspire and inform new types of experimental inquiry. Tool centers develop, maintain, and support analysis and modeling-code repositories. They collect, develop, curate, and implement analysis, modeling, and simulation tools related to GTL tasks and make them available to biology users at GTL centers and in the community. Tool centers would provide a tool repository accessible to investigators or directly by machines in the national grid. For example, users could go to an access point and specify that a tool from a particular repository be used for a task. By coupling activities at GTL facilities and other sites, tool centers would focus on several types of analysis applications:

- Bioinformatics Tools
  - Microbial-community sequence analysis and annotation

  - Proteome and expression-data analysis
- Biophysics Tools
  - Protein dynamics and protein chemistry
  - Protein docking, protein machine modeling and simulation
- Biosystems Tools
  - Metabolic modeling
  - Cell and regulatory-network modeling

Data Centers

Key to GTL's success is genome-scale collection, analysis, dissemination, verification, and modeling of data. Just as with the Human Genome Project and community production of DNA sequence, a key to GTL's success will be the generation of genome-scale data and the data-management and -analysis capabilities needed to interpret the biological "outputs" of a genome. Centrally coordinated data centers, most often located at closely related GTL facilities, would accumulate and integrate data from GTL facilities and distributed projects and organize it for use by the community and GTL modelers. Mirrors of these databases would be supported at biocomputing centers, where the data would be available for incorporation into analysis processes.

Several types of major data resources likely to be needed should be designed in a coordinated way:

- Expression and proteomics databases
- Protein-function and protein-chemistry databases
- Protein-machine, protein-complex, and dynamics database
- Metabolic-pathway and pathway-model database
- Regulatory-network and cell-modeling database
- Microbial-community sequences and annotation database

Biocomputing Centers

The path to understanding the function and dynamic behavior of large molecular systems involves computing, modeling, and simulation of these systems based on structure data, informational parameters, and the use of physically based principles and methods. Biocomputing centers will pool specialized high-performance resources and distributed cluster hardware to provide user access to environments that facilitate large-scale analysis, modeling, and simulation processes. These centers will share a relatively uniform suite of applications (obtained from tool development in GTL facilities and projects) and also mirror databases needed for various analyses or simulations.

While single-molecule simulations currently can be achieved in about a microsecond, the dynamics of protein-protein interactions and simulations of even larger complexes of macromolecules will require much more computing capability. The scale of such simulations poses a significant challenge, requiring capabilities from 50 teraflops to petaflops and beyond (see graphic, p. 31). With a focus on achieving this infrastructure in the next 5 to 10 years, tremendous breakthroughs can be obtained in our understanding of the most fundamentally important macromolecular machines. A similar and somewhat parallel set of requirements will apply to other GTL areas such as network, pathway, and cell modeling.

The infrastructure needed to support computing, modeling, and simulation in GTL facilities and projects will strongly leverage the continuing development of high-performance computing capability within the DOE Advanced Scientific and Computing Research program.

Genomes to Life Program and Facilities Planning Workshops

A series of program-planning workshops has been held to help plan and coordinate Genomes to Life. Meeting reports are placed on the Web as soon as they become available. To learn more about the program, please see the Web site (DOEGenomesToLife.org).

Web site for workshop reports: DOEGenomesToLife.org/pubs.html

2000

October 29–November 1	Genomes to Life Roadmap Planning, San Diego

2001

January 25–27	Genomes to Life Roadmap Planning, Germantown, Md.
June 23	Role of Biotechnology in Mitigating Greenhouse Gas Concentrations, Arlington, Va.
August 7–8	Computational Biology, Germantown, Md.
September 6–7	Computational and Systems Biology, Washington, D.C.
September 9–10	Science Mission Payoffs, Washington, D.C.
October 24–25	Energy and Climate Mission Payoffs, Chicago
December 10–11	Technology Assessment for Mass Spectrometry, Washington, D.C.

2002

January 22–23	Computational Infrastructure, Gaithersburg, Md.
March 6–7	Computer Science, Gaithersburg, Md.
March 18–19	Mathematics, Gaithersburg, Md.
April 16–18	Imaging, Charlotte, N.C.
April 16–19	Computing Strategies, Oak Ridge, Tenn.
June 19–20	Facilities Planning, San Francisco
August 16–17	Facilities Planning, Chicago
October 14–15	Facilities Planning, Gaithersburg, Md.

Entities and Institutions Represented at GTL Workshops and Meetings

Affymetrix • Ames Laboratory • Argonne National Laboratory • Bell Labs • Boston University • Brookhaven National Laboratory • California Institute of Technology • Carnegie Mellon University • Celera Genomics • Columbia University • Cornell University • Dana-Farber Cancer Institute • Duke University • Duke University School of Medicine • DuPont • East Carolina University • Energy Sciences Network (Esnet) • Food and Drug Administration • Genentech Inc. • General Electric • geneticXchange • Harvard Medical School • Harvard University • Hebrew University • InPharmix Inc. • Institute for Systems Biology • IBM • Jefferson Lab • Johns Hopkins School of Medicine • Johns Hopkins University • Joint Genome Institute • Joint Institute for Computational Science • Keck Graduate Institute • Keio University • Lawrence Berkeley National Laboratory • Lawrence Livermore National Laboratory • Los Alamos National Laboratory • Marshfield Medical Research Foundation • Massachusetts Institute of Technology • Medical University of South Carolina • Merck Research Laboratories • Molecular Sciences Institute • Monsanto Company • Montana State University at Bozeman • Monterey Bay Aquarium Research Institute • National Academy of Sciences • National Cancer Institute • National Center for Biotechnology Information • National Center for Genome Research • National Center for Supercomputing Applications • National Energy Research Scientific Computing Center • National Human Genome Research Institute • National Institute of General Medical Sciences • National Institutes of Health • National Renewable Energy Laboratory • National Research Council • National Science Foundation • National Water Research Institute • Natural Resources Defense Council • New England Complex Systems Institute • New York University • North Carolina Supercomputing Center • Novation Biosciences • Oak Ridge Institute for Science and Education • Oak Ridge National Laboratory • Office of Management and Budget • Ohio State University • Pacific Northwest National Laboratory • Pittsburgh Supercomputing Center • Princeton University • Rockefeller University • Sandia National Laboratories • Sanger Centre • Scripps Institution of Oceanography • Scripps Research Institute • Southwest Parallel Software • SRI International • Stanford School of Medicine • Stanford University • Test Measurement Systems Inc. • Texas Tech University • The Institute for Genomic Research • The Packson Laboratory • United State Department of Agriculture • University of California, Berkeley • University of California, Irvine • University of California, Los Angeles • University of California, San Diego • University of California, San Francisco • University of California, Santa Barbara • University of Colorado • University of Connecticut Health Center • University of Florida • University of Illinois • University of Illinois at Urbana-Champaign • University of Iowa • University of Maryland Biotechnology Institute • University of Massachusetts, Amherst • University of Miami • University of Michigan • University of Pennsylvania • University of Pittsburgh • University of Southern California • University of Texas • University of Utah • University of Washington • University of Wisconsin • Vanderbilt University • Vertex Pharmaceuticals Inc. • Weyerhauser Company • Whitehead Institute for Genome Research

U.S. Department of Energy Office of Science

Marvin Frazier
Office of Biological and Environmental Research (SC-72)
301/903-5468, Fax: 301/903-8521
marvin.frazier@science.doe.gov

Gary Johnson
Office of Advanced Scientific Computing Research (SC-30)
301/903-5800, Fax: 301/903-7774
garyj@er.doe.gov